California Housing Cost Burden Analysis Notebook

Introduction

This purpose of this project is to analyze the California housing cost burden between the years 2006 and 2010 in order to explore any potential patterns, trends and correlations within this time span. Additionally, from this analysis we can identify and quantify to an extent which counties/cities/demographics in California suffer the most these cost burdens.

Motivation / Context

My personal motivation for performing this analysis was to learn more about how affordable (or unaffordable) the housing market was in California around 10-15 years ago, the severity of the cost burdens placed on families California as well as what characterized the types of communities that were most disadvantaged by this. Today, it is commonly known that the housing costs in California (especially in the Bay Area where I live) are extremely high yet competitive among buyers. As someone who has yet to own his first home, understanding what the housing cost burden has been historically compared to how it is now, can help to forecast the direction of the housing market as well as help me to set reasonable expectation in the future, should I choose to become a California home owner.

The dataset used for this analysis captures a variety of information pertaining to housing and California residents. This includes economic data such as income level of household and the percentage of households paying more than 30% (or 50%) of their monthly household income towards housing costs. In additional, geographical housing data such as geographic type, region name and region code, as well as, other demographic data including racial/ethnic group, are all shown on our dataset.

Limitations

As stated from the source material:

"The housing cost burden estimates do not adjust for differences in household size. Estimates for the survey period 2006-2010 are bisected by the Great Recession (2008), marked by a large increase in home foreclosures, and house/rental price instability. Due to changes in definitions and sampling, HUD does not recommend making comparisons to prior years' estimates. ACS data are available at census tract geographies, albeit with a definition of cost burden that is different than that of CHAS."

Data Preparation & Cleaning

The housing cost dataset used for this analysis was downloaded from the California Data Portal linked below:

[https://data.ca.gov/dataset/housing-cost-burden (https://data.ca.gov/dataset/housing-cost-burden)](https://data.ca.gov/dataset/housing-cost-burden)

The dataset is downloaded as a .xlsx file by default. We convert this to a standard .csv file on Excel. Upon inspection of our .csv file on MS Excel, we see that there are 521265 total observations (rows) and 26 attributes (columns).

We will analyze this data using the Pandas Dataframe.

In [1]:
```
# Import libraries

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from matplotlib import rc
import matplotlib.patches as mpatches
import seaborn as sns
import plotly.figure_factory as ff
import plotly.express as px
from chart_studio import plotly
import plotly.graph_objects as go
from plotly.subplots import make_subplots
import requests
import json
```

In [2]:
```
data = pd.read_csv(r"/Users/julian/Documents/Work/Data Analytics/Data Analytics Portfolio/CA Housing Cost Burden/Raw data/hci_acs_chas_raceincome_housingcostburden_ct_pl_co_re_st_7-30-14-ada.csv")
```

```
/opt/anaconda3/lib/python3.7/site-packages/IPython/core/interactiveshell.py:3058: DtypeWarning: Columns (0,11) have mixed types. Specify dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

In [3]:
```python
# Preview our imported CA Housing Cost Burden data
data.head()
```

Out[3]:

	ind_id	ind_definition	datasource	reportyear	burden	tenure	race_eth_code	race_e
0	106	Percent of households spending more than 30% (...	CHAS	2006-2010	> 30% of monthly household income consumed by ...	Owner-occupied households	9.0	
1	106	Percent of households spending more than 30% (...	CHAS	2006-2010	> 30% of monthly household income consumed by ...	Owner-occupied households	9.0	
2	106	Percent of households spending more than 30% (...	CHAS	2006-2010	> 50% of monthly household income consumed by ...	Owner-occupied households	9.0	
3	106	Percent of households spending more than 30% (...	CHAS	2006-2010	> 50% of monthly household income consumed by ...	Owner-occupied households	9.0	
4	106	Percent of households spending more than 30% (...	CHAS	2006-2010	> 30% of monthly household income consumed by ...	Renter-occupied households	9.0	

5 rows × 26 columns

From simply calling our data, we verify that the import into Python was successful as it appears as a table of 521265 rows × 26 columns

Not all the columns that we've imported into Python will probably be useful for our analysis. Therefore, we should further understand what each of the columns represent in close detail and then remove the columns we don't need.

Looking at the "Data Dictionary" file that we also downloaded from the California Data Portal URL, each column column, definition and format is shown on the table below. (For coding information, see original .csv or xlsx file since its too much to display here)

Column Name	Definition	Format
ind_id	Indicator ID	Plain Text
ind_definition	Definition of indicator in plain language	Plain Text
datasource	Source of the indicator data	Plain Text
reportyear	Year(s) that the indicator was reported	Plain Text
burden	Description of housing cost burden strata	Plain Text
tenure	Description of housing tenure	Plain Text
race_eth_code	numeric code for a race/ethnicity group	Plain Text
race_eth_name	Name of race/ethnic group	Plain Text
income_level	Income level (n=3)	Plain Text
geotype	Type of geographic unit	Plain Text
geotypevalue	Value of geographic unit	Plain Text
geoname	Name of geographic unit	Plain Text
county_name	Name of county that geotype is in	Plain Text
county_fips	FIPS code of county that geotype is in	Plain Text
region_name	Metropolitan Planning Organization (MPO) - based region name	Plain Text
region_code	Metropolitan Planning Organization (MPO) - based region code	Plain Text
total_households	Number of owner- and renter - occupied households; the denominator for this indicator	Numeric
burdened_households	Number of households carrying a > 30% (>50%) housing cost burden; the numerator for this indicator	Numeric
percent	Percent of households carrying a > 30% (>50%) housing cost burden; the numerator for this indicator	Numeric
LL95CI	Lower Limit of 95% confidence interval	Numeric
UL95CI	Upper Limit of 95% confidence interval	Numeric
SE	Standard error of percent	Numeric
rse	Relative standard error (se/percent * 100) expressed as a percent	Numeric
CA_decile	California decile	Numeric
CA_RR	Rate ratio to California rate	Numeric
version	Date/time stamp of version of data	Date/Time

Based on review of the Data Dictionary summary table, the data columns that don't appear to be very useful for the scope of this analysis and can be deleted. These particular columns and rationale for deletion is shown in the table below:

Column Name	Rationale for Deletion
ind_id	Not needed since all values are the same (106)
datasource	Not used in this analysis as datasource can easily be retrieved from raw file
reportyear	Not needed since all values are the same (2006 - 2010)
region_code	No added information that "region_name" column does not already provide
race_eth_code	No added information that "race_eth_name" column does not already provide
geoname	Information not necessary for this analysis
LL95CI	Not needed since "rse" will be used to assess data reliability
UL95CI	Not needed since "rse" will be used to assess data reliability
SE	Not needed since "rse" will be used to assess data reliability
CA_decile	Information not necessary for this analysis
CA_RR	Information not necessary for this analysis
version	Not needed since all values are the same (29June2014)

In [4]:

```
# We shall remove the unneeded columns we've identified below with the following lines of code
data_mod1 = data.drop(['ind_id', 'datasource','reportyear', 'region_code', 'race_eth_code','geoname', 'LL95CI', 'UL95CI', 'SE', 'CA_decile', 'CA_RR', 'version'], axis=1);

# Preview and confirm change
data_mod1.head()
```

Out[4]:

	ind_definition	burden	tenure	race_eth_name	income_level	geotype	geotypevalue
0	Percent of households spending more than 30% (...	> 30% of monthly household income consumed by ...	Owner-occupied households	Total	Monthly household income at <=30% of HUD-adjus...	CA	6.0
1	Percent of households spending more than 30% (...	> 30% of monthly household income consumed by ...	Owner-occupied households	Total	Monthly household income at all levels of HUD-...	CA	6.0
2	Percent of households spending more than 30% (...	> 50% of monthly household income consumed by ...	Owner-occupied households	Total	Monthly household income at <=30% of HUD-adjus...	CA	6.0
3	Percent of households spending more than 30% (...	> 50% of monthly household income consumed by ...	Owner-occupied households	Total	Monthly household income at all levels of HUD-...	CA	6.0
4	Percent of households spending more than 30% (...	> 30% of monthly household income consumed by ...	Renter-occupied households	Total	Monthly household income at <=30% of HUD-adjus...	CA	6.0

In [5]:
```
# Now, we'll perform a count for all the values in the remaining colum
ns.
data_mod1.count()
```

Out[5]:
```
ind_definition          521262
burden                  521262
tenure                  521262
race_eth_name           521262
income_level            521262
geotype                 521262
geotypevalue            521262
county_name             520452
county_fips             520452
region_name             521208
total_households        213180
burdened_households     213180
percent                 179994
rse                     158508
dtype: int64
```

In [6]:
```
# Then, perform a null count.
data_mod1.isnull().sum()

# Something important to note is that a value of "0" is typically trea
ted as null. Therefore we have to verify that a null/zero
# were appropriate given the corresponding column values in the approp
riate context
```

Out[6]:
```
ind_definition               3
burden                       3
tenure                       3
race_eth_name                3
income_level                 3
geotype                      3
geotypevalue                 3
county_name                813
county_fips                813
region_name                 57
total_households        308085
burdened_households     308085
percent                 341271
rse                     362757
dtype: int64
```

In [7]:
```
# Right off the bat, we recognize a data cleaning opportunity based on our outputs above.
# There are 3 rows that are null across all of our columns that we can remove from dataframe.
# We'll perform this deletion based off of our "ind_definition" column
data_mod2 = data_mod1.dropna(subset = ["ind_definition"], inplace=False);

# Verify deletion
data_mod2.isnull().sum()
```

Out[7]:
```
ind_definition              0
burden                      0
tenure                      0
race_eth_name               0
income_level                0
geotype                     0
geotypevalue                0
county_name               810
county_fips               810
region_name                54
total_households       308082
burdened_households    308082
percent                341268
rse                    362754
dtype: int64
```

In [8]:
```
# We've successfully removed the null rows pertaining to those first 7 columns.
# Now let's validate that the 810 null values for our "county" columns is valid.
# The columns "county_name" and "county_fips" should be null when the "geotype" value is "CA" or "RE".
# In other words, there isn't a county name for observations that are state-wide or region-wide.

# Count the number of rows where "geotype" is CA or RE and "county_name" is null.
len(data_mod2[((data_mod2['geotype'] == 'CA') | (data_mod2['geotype'] == 'RE')) & (data_mod2['county_name'].isna())])

# If the result is 810 (as expected), then all of the nulls for "county_name" are valid.
```

Out[8]: 810

In [9]:
```
# We repeat the same process for "county_fips".
# Count the number of rows where "geotype" is CA or RE and "county_fip
s" is null.
len(data_mod2[((data_mod2['geotype'] == 'CA') | (data_mod2['geotype']
== 'RE')) & (data_mod2['county_fips'].isna())])

# If the result is 810 (as expected), then all of the nulls for "count
y_fips" are valid.
```
Out[9]: 810

In [10]:
```
# Now we'll address the 54 null values for "region_name".
# Of all the "geotype" values, only CA should not have an assigned reg
ion value.
# This is because a state-wide observation cannot be classifed or assi
gned to particular region,
# whereas CO, CT, PL, RE all fall within a region.

# Count the number of rows where "geotype" is CA and "county_name" is
null.
len(data_mod2[(data_mod2['geotype'] == 'CA') & (data_mod2['region_name
'].isna())])

# If the result is 54 (as expected), then all of the nulls for "region
_name" are valid.
```
Out[10]: 54

In [11]:
```
# We now shift our attention to the 308082 null values that were ident
ifed for the columns "total_households" and
# "burdened_households". We can assume but will verify that these obse
rvations simply represent places where no one
# of that particular demographic, burden level or etc. is represented
in that location

# Count the number of cases where "total_households" is null but "burd
ened_households" is not null.
Hou_NotZero = data_mod2['total_households'].isna() & data_mod2['burden
ed_households'].notna()
Hou_NotZero.value_counts()


# Count the number of rows where both "total_households" and "burdened
_households" are null.
len(data_mod2[(data_mod2['total_households'].isna()) & (data_mod2['bur
dened_households'].isna())])

# If the result is 308082 (as expected), then we've successfully verif
ied that no burdened households were mistakenly
# counted in observations that had zero corresponding households as th
at would be an impossible situation and
# clearly an error. However, the opposite situation where we have no b
urdened households despite a non-zero total
# number of households is a possible scenario.
```

Out[11]: 308082

In [12]:
```
# For the "percent" column, there are 341268 null values. Practically
speaking, the only situation where percent is null
# null would be if and only if the corresponding "burdened_households"
value was also null. In other words, the number
# of "burdened households" being null/zero should be the sole instance
that translates to a null "percent"
# We perform our check again to ensure the inverse of this is always F
alse.

# Count the number of rows where both "percent" and "burdened_househol
ds" are null.
len(data_mod2[(data_mod2['percent'].isna()) & (data_mod2['burdened_hou
seholds'].isna())])

# The output only shows a total count of 308082, which shows there is
another for "burdened_households", that yields
# a null "percent"
```

Out[12]: 308082

In [13]:
```
# As mentioned earlier, the value 0 can sometimes not be counted as a null. Therefore, we'll include a check that
# validates the instance where "burdened_households" is 0 as well.

# Count the number of rows where "percent" is null and "burdened_households" is 0.
len(data_mod2[(data_mod2['percent'].isna()) & (data_mod2['burdened_households'] == 0)])

# If the sum of this output and the previous output (308082) equals 341268,  then all of the nulls for "percent"
# are valid.
```

Out[13]: 33186

In [14]:
```
# Now that we know there is no invalid cases where percent is incorrectly null (or zero), we continue the data cleaning
# process by looking at "rse", which represents the Relative Standard Error. This value is an indicator of how reliabile
# the data is based off of the sample size taken. A null (or zero) value in this context is actually a good sign.
# However, based on the Data Dictionary file, an "rse" of 23 percent or more means that there was not a sufficient
# sample size taken and the data is considered unreliable.
# For this final step of the data cleaning process, we remove all rows where "rse" >= 23.

data_mod3 = data_mod2.drop(data_mod2[data_mod2.rse >= 23].index, inplace = False)

# Verify that there are no rows where "rse" >= 23
data_mod3[data_mod3.rse >= 23].shape[0]

# Once we've verified our rows have acceptable "rse" values, we can drop the column from our df since its no longer needed
data_mod4 = data_mod3.drop(['rse'], axis = 1)

# This concludes our data cleaning process check for any erroneous and unwanted null values in our dataframe.
```

In [15]:
```
# Next, let's verify the data types of our dataframe.
data_mod4.dtypes

# Comparing the output of our data types to what was defined in the da
ta dictionary, makes sense in the context of this analysis.
# No datatype conversions are needed at this time.
```

Out[15]:
```
ind_definition           object
burden                   object
tenure                   object
race_eth_name            object
income_level             object
geotype                  object
geotypevalue            float64
county_name              object
county_fips             float64
region_name              object
total_households        float64
burdened_households     float64
percent                 float64
dtype: object
```

In [16]:
```
# We notice that "geotypevalue" is a float64. Per the data dictionary
and depending on the indicator context,
# this column represents either an 11 digit FIPS census tract code, a
5 digit FIPS code (place or county) or 2
# digit FIPS code (region or state). As we noticed from the original o
utput preview using the "heads" function,
# All the values in the preview came out as "6.0" due to it being a fl
oating value. For the context described,
# it is appropriate that we convert this value into an integer as foll
ows.

data_mod4['geotypevalue'] = data_mod4['geotypevalue'].astype('int')
```

In [17]:
```
# Now,we'll re-verify the change to the geotypevalue data type.
data_mod4.dtypes
```

Out[17]:
```
ind_definition          object
burden                  object
tenure                  object
race_eth_name           object
income_level            object
geotype                 object
geotypevalue             int64
county_name             object
county_fips            float64
region_name             object
total_households       float64
burdened_households    float64
percent                float64
dtype: object
```

In [18]:
```
# And lastly, confirm all the values are properly listed using the "he
ad" function to preview
print(data_mod4.geotypevalue)
```

```
0             6
1             6
2             6
3             6
4             6
          ...
521259    86804
521260    86804
521261    86804
521262    86804
521263    86804
Name: geotypevalue, Length: 444479, dtype: int64
```

Analysis

Now that the data preparation and cleaning process is complete,we can proceed with creating subsets of interest for our dataframe and performing our exploratory analysis. Since our table captures the cost burden indicators based on various attributes of the study (ethnicity, tenure, income leve, etc) creating ad-hoc subsets will assist us in organizing and eventually visualizing our data.

In this section, we will specifically aim to answer the following questions:

1. In all of California (by county), what percentage of households are affected by the following cost burdens, broken down by region in descending order? What is the mean and median for each?

```
a. Burden > 30% (gross rent + selected housing cost), All HUD-adjusted
income levels & ethnic groups
b. Burden > 50% (gross rent + selected housing cost), All HUD-adjusted
income levels & ethnic groups
```

2. In the Bay Area Region (by county), what percentage of households are affected the following cost burdens?

```
a. Burden > 30% (gross rent + selected housing cost), All HUD-adjusted
income levels & ethnic groups
b. Burden > 50% (gross rent + selected housing cost), All HUD-adjusted
income levels & ethnic groups
```

3. For all of California (by race/ethnic group), what percentage of households are affected the following cost burdens?

```
a. Burden > 30% (gross rent + selected housing costs), Owner-occupied
Tenure, All income levels
b. Burden > 50% (gross rent + selected housing costs), Owner-occupied
Tenure, All income levels
c. Burden > 30% (gross rent), Renter-occupied Tenure, All income level
s
d. Burden > 50% (gross rent), Renter-occupied Tenure, All income level
s
```

4. In all of California (by CT), what percentage of households (all race/ethnic groups) are affected by a cost burden >= 50% monthly household income? What is the mean and median for each?

```
a. Mortgage paying, owner occupied households
b. Rent paying, renter occupied households
```

5. In the Bay Area Region (by county), what percentage of households (all race/ethnic groups) are affected by a cost burden >= 50% monthly household income? What is the mean and median for each?

```
a. Mortgage paying, owner occupied households
b. Rent paying, renter occupied households
```

Question 1a

In [19]:
```
# We will create a subset of our dataframe that captures the rows of i
nterest as specifed by 1a.
# 1. In all of California (by county), what percentage of households a
re affected by the following cost burdens, broken down by region in de
scending order? What is the mean and median for each?
# What is the mean and median for each?
#      a. Burden > 30% (gross rent + selected housing cost), All HUD-a
djusted income levels & ethnic groups

df_1a = data_mod4[(data_mod4.geotype == 'CO') & (data_mod4.burden == '
> 30% of monthly household income consumed by monthly, gross rent or s
elected housing costs') & (data_mod4.income_level == 'Monthly househol
d income at all levels of HUD-adjusted family median income') & (data_
mod4.race_eth_name == 'Total')];
```

In [20]:
```
# Show the 5 counties (state-wide) with the Highest Percentage of Hous
eholds with the specified burden criteria
df_1a_top5 = df_1a.sort_values(by='percent',ascending = False)
df_1a_top5 = df_1a_top5.head(5)
df_1a_top5 = df_1a_top5[['county_name', 'total_households', 'burdened_
households', 'percent']]
print(df_1a_top5)
```

```
        county_name  total_households  burdened_households    percen
t
248518   San Benito           16810.0               8370.0  49.79179
1
1791    Los Angeles         3217890.0            1552720.0  48.25273
7
248410    Riverside          666905.0             318985.0  47.83065
1
249004   Santa Cruz           93800.0              44605.0  47.55330
5
2169           Mono            5285.0               2460.0  46.54683
1
```

In [21]:
```
# Show the 5 counties (state-wide) with the Lowest Percentage of House
holds with the specified burden criteria
df_1a_bottom5 = df_1a.sort_values(by ='percent', ascending = True)
df_1a_bottom5 = df_1a_bottom5.head(5)
df_1a_bottom5 = df_1a_bottom5[['region_name', 'total_households', 'bur
dened_households', 'percent']]
print(df_1a_bottom5)
```

```
                        region_name  total_households  burdened_househo
lds  \
494165                  North Coast            5890.0                168
5.0
2115               Northeast Sierra            3975.0                118
4.0
1953    Central/Southeast Sierra            7725.0                247
5.0
1521    Central/Southeast Sierra            7980.0                274
5.0
1629             San Joaquin Valley           40605.0               1420
5.0

          percent
494165  28.607810
2115    29.786164
1953    32.038835
1521    34.398496
1629    34.983376
```

In [22]:
```
# Calculate mean
df_1a.percent.mean()
```

Out[22]: 41.833537853928554

In [23]:
```
# Calculate median
df_1a.percent.median()
```

Out[23]: 42.47537756

Question 1b

In [24]:
```
# We will create a subset of our dataframe that captures the rows of i
nterest as specifed by 1b.
# 1. In all of California (by county), what percentage of households a
re affected by the following cost burdens, broken down by region in de
scending order?
# What is the mean and median for each?
#      b. Burden > 50% (gross rent + selected housing cost), All HUD-a
djusted income levels & ethnic groups

df_1b = data_mod4[(data_mod4.geotype == 'CO') & (data_mod4.burden == '
> 50% of monthly household income consumed by monthly, gross rent or s
elected housing costs') & (data_mod4.income_level == 'Monthly househol
d income at all levels of HUD-adjusted family median income') & (data_
mod4.race_eth_name == 'Total')];
```

In [25]:
```
# Show the 5 counties (state-wide) with the Highest Percentage of Hous
eholds with the specified burden criteria
df_1b_top5 = df_1b.sort_values(by='percent',ascending = False)
df_1b_top5 = df_1b_top5.head(5)
df_1b_top5 = df_1b_top5[['county_name', 'total_households', 'burdened_
households', 'percent']]
print(df_1b_top5)
```

```
        county_name  total_households  burdened_households    percen
t
1793    Los Angeles         3217890.0             787845.0  24.48327
9
248520   San Benito           16810.0               4020.0  23.91433
7
2009      Mendocino           34375.0               8220.0  23.91272
7
248412    Riverside          666905.0             151320.0  22.68988
8
249006   Santa Cruz           93800.0              20955.0  22.34008
5
```

In [26]:
```
# Show the 5 counties (state-wide) with the Lowest Percentage of House
holds with the specified burden criteria
df_1b_bottom5 = df_1b.sort_values(by ='percent', ascending = True)
df_1b_bottom5 = df_1b_bottom5.head(5)
df_1b_bottom5 = df_1b_bottom5[['county_name', 'total_households', 'bur
dened_households', 'percent']]
print(df_1b_bottom5)
```

```
        county_name  total_households  burdened_households    percent
494167      Trinity            5890.0                625.0  10.611205
1091         Colusa            6970.0                955.0  13.701578
1955       Mariposa            7725.0               1145.0  14.822006
1631          Kings           40605.0               6145.0  15.133604
1523           Inyo            7980.0               1270.0  15.914787
```

In [27]:
```
# Calculate mean
df_1b.percent.mean()
```

Out[27]: 19.51751523690909

In [28]:
```
# Calculate median
df_1b.percent.median()
```

Out[28]: 20.023739300000003

Question 2a

In [29]:
```
# We will create a subset of our dataframe that captures the rows of i
nterest as specifed by 2a.
# 2. In the Bay Area Region (by county), what percentage of households
are affected the following cost burdens?
#       a. Burden > 30% (gross rent + selected housing cost), All HUD-
adjusted income levels & ethnic groups

df_2a = data_mod4[(data_mod4.geotype == 'CO') & (data_mod4.region_name
== 'Bay Area') & (data_mod4.burden == '> 30% of monthly household inco
me consumed by monthly, gross rent or selected housing costs') & (data
_mod4.income_level == 'Monthly household income at all levels of HUD-a
djusted family median income') & (data_mod4.race_eth_name == 'Total')]
;
```

In [30]:
```
# Show Results of Counties with the Percentage of Households (desc) wi
th the specified burden criteria
df_2a_all = df_2a.sort_values(by='percent',ascending = False)
df_2a_all = df_2a_all[['county_name', 'total_households', 'burdened_ho
useholds', 'percent']]
print(df_2a_all)
```

```
            county_name  total_households  burdened_households    perc
ent
249220           Solano          139010.0              62735.0  45.129
847
249274           Sonoma          184035.0              82905.0  45.048
496
1143       Contra Costa          368085.0             165515.0  44.966
516
1899              Marin          102725.0              45305.0  44.103
188
819             Alameda          532025.0             233325.0  43.856
022
2277               Napa           49180.0              20990.0  42.679
951
248842        San Mateo          255760.0             105540.0  41.265
249
248950      Santa Clara          596745.0             240480.0  40.298
620
248680    San Francisco          335955.0             132510.0  39.442
783
```

Question 2b

In [31]:
```
# We will create a subset of our dataframe that captures the rows of i
nterest as specifed by 2b.
# 2. In the Bay Area Region (by county), what percentage of households
are affected the following cost burdens?
#       b. Burden > 50% (gross rent + selected housing cost), All HUD-
adjusted income levels & ethnic groups

df_2b = data_mod4[(data_mod4.geotype == 'CO') & (data_mod4.region_name
== 'Bay Area') & (data_mod4.burden == '> 50% of monthly household inco
me consumed by monthly, gross rent or selected housing costs') & (data
_mod4.income_level == 'Monthly household income at all levels of HUD-a
djusted family median income') & (data_mod4.race_eth_name == 'Total')]
;
```

In [32]:
```
# Show Results of Counties with the Percentage of Households (desc) wi
th the specified burden criteria
df_2b_all = df_2b.sort_values(by='percent',ascending = False)
df_2b_all = df_2b_all[['county_name', 'total_households', 'burdened_ho
useholds', 'percent']]
print(df_2b_all)
```

```
          county_name  total_households  burdened_households    perc
ent
1901            Marin          102725.0              22095.0  21.508
883
821           Alameda          532025.0             111415.0  20.941
685
249276         Sonoma          184035.0              38325.0  20.824
843
1145     Contra Costa          368085.0              74310.0  20.188
272
249222         Solano          139010.0              27835.0  20.023
739
2279             Napa           49180.0               9755.0  19.835
299
248682  San Francisco          335955.0              64545.0  19.212
395
248844      San Mateo          255760.0              48585.0  18.996
325
248952    Santa Clara          596745.0             109945.0  18.424
117
```

Question 3a

In [33]:
```
# We will create a subset of our dataframe that captures the rows of i
nterest as specifed by 3a.
# 3. For all of California (by race/ethnic group), what percentage of
households are affected the following cost burdens?
#       a. Burden > 30% (gross rent + selected housing costs), Owner-o
ccupied household, All income levels

df_3a = data_mod4[(data_mod4.geotype == 'CA') & (data_mod4.burden == '
> 30% of monthly household income consumed by monthly, selected, housi
ng costs') & (data_mod4.tenure == 'Owner-occupied households') & (data
_mod4.income_level == 'All income levels')];
```

In [34]:
```
# Show the Percentages of Households with the specified burden criteri
a by race/ethnic group
df_3a_all = df_3a.sort_values(by='percent',ascending = False)
df_3a_all = df_3a_all[['race_eth_name', 'total_households', 'burdened_
households', 'percent']]
print(df_3a_all)
```

```
   race_eth_name  total_households  burdened_households    percent
30        Latino         1533770.0             794040.0  51.770474
24     AfricanAm          310565.0             160380.0  51.641363
36         NHOPI           17320.0               8740.0  50.461894
48      Multiple          111985.0              50360.0  44.970309
18         Asian          853465.0             370415.0  43.401311
12          AIAN           28855.0              11415.0  39.559868
42         White         4256085.0            1534255.0  36.048505
```

Question 3b

In [35]:
```
# We will create a subset of our dataframe that captures the rows of i
nterest as specifed by 3b.
# 3. For all of California (by race/ethnic group), what percentage of
households are affected the following cost burdens?
#       b. Burden > 50% (gross rent + selected housing costs), Owner-o
ccupied Tenure, All income levels

df_3b = data_mod4[(data_mod4.geotype == 'CA') & (data_mod4.burden == '
> 50% of monthly household income consumed by monthly, selected, housi
ng costs') & (data_mod4.tenure == 'Owner-occupied households') & (data
_mod4.income_level == 'All income levels')];
```

In [36]:
```
# Show the Percentages of Households with the specified burden criteri
a by race/ethnic group
df_3b_all = df_3b.sort_values(by='percent',ascending = False)
df_3b_all = df_3b_all[['race_eth_name', 'total_households', 'burdened_
households', 'percent']]
print(df_3b_all)
```

```
   race_eth_name  total_households  burdened_households    percent
31        Latino         1533770.0             387870.0  25.288668
25     AfricanAm          310565.0              77390.0  24.919099
37         NHOPI           17320.0               3985.0  23.008083
49      Multiple          111985.0              21845.0  19.507077
19         Asian          853465.0             163410.0  19.146655
13          AIAN           28855.0               5175.0  17.934500
43         White         4256085.0             642600.0  15.098383
```

Question 3c

In [37]:
```
# We will create a subset of our dataframe that captures the rows of i
nterest as specifed by 3c.
# 3. For all of California (by race/ethnic group), what percentage of
households are affected the following cost burdens?
#       c. Burden > 30% (gross rent + selected housing costs), Renter-
occupied Tenure, All income levels

df_3c = data_mod4[(data_mod4.geotype == 'CA') & (data_mod4.burden == '
> 30% of monthly household income consumed by monthly, gross rent') &
(data_mod4.tenure == 'Renter-occupied households') & (data_mod4.income
_level == 'All income levels')];
```

In [38]:
```
# Show the Percentages of Households with the specified burden criteri
a by race/ethnic group
df_3c_all = df_3c.sort_values(by='percent',ascending = False)
df_3c_all = df_3c_all[['race_eth_name', 'total_households', 'burdened_
households', 'percent']]
print(df_3c_all)
```

```
   race_eth_name  total_households  burdened_households   percent
26     AfricanAm          491350.0             284510.0  57.903735
32        Latino         1770415.0             982630.0  55.502806
14          AIAN           28595.0              14925.0  52.194440
50      Multiple          117615.0              57770.0  49.117885
44         White         2231375.0            1039455.0  46.583609
38         NHOPI           18930.0               8790.0  46.434231
20         Asian          622525.0             273240.0  43.892213
```

Question 3d

In [39]:
```
# We will create a subset of our dataframe that captures the rows of i
nterest as specifed by 3d.
# 3. For all of California (by race/ethnic group), what percentage of
households are affected the following cost burdens?
#       d. Burden > 50% (gross rent + selected housing costs), Renter-
occupied Tenure, All income levels

df_3d = data_mod4[(data_mod4.geotype == 'CA') & (data_mod4.burden == '
> 50% of monthly household income consumed by monthly, gross rent') &
(data_mod4.tenure == 'Renter-occupied households') & (data_mod4.income
_level == 'All income levels')];
```

In [40]:
```
# Show the Percentages of Households with the specified burden criteri
a by race/ethnic group
df_3d_all = df_3d.sort_values(by='percent',ascending = False)
df_3d_all = df_3d_all[['race_eth_name', 'total_households', 'burdened_
households', 'percent']]
print(df_3d_all)
```

```
   race_eth_name  total_households  burdened_households    percent
27     AfricanAm          491350.0             160505.0  32.666124
15          AIAN           28595.0               8110.0  28.361602
33        Latino         1770415.0             498290.0  28.145378
51      Multiple          117615.0              30710.0  26.110615
45         White         2231375.0             527825.0  23.654697
21         Asian          622525.0             141005.0  22.650496
39         NHOPI           18930.0               4090.0  21.605917
```

Question 4a

In [41]:
```
# We will create a subset of our dataframe that captures the rows of i
nterest as specifed by 4a.
# 4. In all of California (by CT), what percentage of households (all
race/ethnic groups) are affected by a cost burden >= 50% monthly house
hold income?
# What is the mean and median for each?
#       a. Mortgage paying, owner occupied households

df_4a = data_mod4[(data_mod4.geotype == 'CT') & (data_mod4.burden == '
>= 50% of monthly household income consumed by monthly, selected housi
ng costs') & (data_mod4.race_eth_name == 'Total') & (data_mod4.tenure
== 'Mortgage-paying, owner-occupied households')];
df_4a = df_4a.dropna(subset = ["percent"], inplace=False);
```

In [42]:
```
# Show the 5 Census Tracts (minimum of 50 households) with the Highest Percentage of Households with the specified burden criteria
df_4a_top5 = df_4a.sort_values(by='percent',ascending = False)
df_4a_top5 = df_4a_top5[df_4a_top5["total_households"]>=50]
df_4a_top5 = df_4a_top5.head(5)
df_4a_top5 = df_4a_top5[['county_name', 'geotypevalue', 'total_households', 'burdened_households', 'percent']]
print(df_4a_top5)
```

```
           county_name  geotypevalue  total_households  burdened_hou
seholds  \
92276      Los Angeles    6037235201             512.0
305.0
279981  San Bernardino    6071003509             596.0
348.0
186776     Los Angeles    6037600202             405.0
233.0
166958     Los Angeles    6037532900             246.0
139.0
195848     Los Angeles    6037700600            1049.0
591.0

          percent
92276   59.570312
279981  58.389262
186776  57.530864
166958  56.504065
195848  56.339371
```

In [43]:
```
# Show the 5 Census Tracts (minimum of 50 households) with the Lowest
Percentage of Households with the specified burden criteria
df_4a_bottom5 = df_4a.sort_values(by ='percent', ascending = True)
df_4a_bottom5 = df_4a_bottom5[df_4a_bottom5["total_households"]>=50]
df_4a_bottom5 = df_4a_bottom5.head(5)
df_4a_bottom5 = df_4a_bottom5[['county_name', 'geotypevalue', 'total_h
ouseholds', 'burdened_households', 'percent']]
print(df_4a_bottom5)
```

```
           county_name  geotypevalue  total_households  burdened_hou
seholds  \
320103  San Bernardino    6071010422             170.0
0.0
317880          Shasta    6089010100              65.0
0.0
316908       San Diego    6073010013              78.0
0.0
312759       San Diego    6073009509             396.0
0.0
59876      Los Angeles    6037125320              61.0
0.0

        percent
320103      0.0
317880      0.0
316908      0.0
312759      0.0
59876       0.0
```

In [44]:
```
# Calculate mean
df_4a.percent.mean()
```

Out[44]: 23.43436070339273

In [45]:
```
# Calculate median
df_4a.percent.median()
```

Out[45]: 22.502446185

Question 4b

In [46]:
```
# We will create a subset of our dataframe that captures the rows of i
nterest as specifed by 4b.
# 4. In all of California (by CT), what percentage of households (all
race/ethnic groups) are affected by a cost burden >= 50% monthly house
hold income? What is the mean and median for each?
#      b. Rent paying, renter occupied households

df_4b = data_mod4[(data_mod4.geotype == 'CT') & (data_mod4.burden == '
>= 50% of monthly household income consumed by monthly, gross rent')
& (data_mod4.race_eth_name == 'Total') & (data_mod4.tenure == 'Rent-pa
ying, renter-occupied households')];
df_4b = df_4b.dropna(subset = ["percent"], inplace=False);
```

In [47]:
```
# Show the 5 Census Tracts (minimum of 50 households) with the Highest
Percentage of Households with the specified burden criteria
df_4b_top5 = df_4b.sort_values(by='percent',ascending = False)
df_4b_top5 = df_4b_top5[df_4b_top5["total_households"]>=50]
df_4b_top5 = df_4b_top5.head(5)
df_4b_top5 = df_4b_top5[['county_name', 'geotypevalue', 'total_househo
lds', 'burdened_households', 'percent']]
print(df_4b_top5)
```

```
            county_name  geotypevalue  total_households  burdened_ho
useholds  \
322272  San Luis Obispo    6079010902            1462.0
1093.0
97253       Los Angeles    6037265304            1026.0
750.0
322164  San Luis Obispo    6079010901            1087.0
712.0
13873            Fresno    6019001000             540.0
342.0
323460  San Luis Obispo    6079011200            1612.0
1005.0

          percent
322272  74.760602
97253   73.099415
322164  65.501380
13873   63.333333
323460  62.344913
```

In [48]:
```
# Show the 5 Census Tracts (minimum of 50 households) with the Lowest Percentage of Households with the specified burden criteria
df_4b_bottom5 = df_4b.sort_values(by='percent',ascending = True)
df_4b_bottom5 = df_4b_bottom5[df_4b_bottom5["total_households"]>=50]
df_4b_bottom5 = df_4b_bottom5.head(5)
df_4b_bottom5 = df_4b_bottom5[['county_name', 'geotypevalue','total_households', 'burdened_households', 'percent']]
print(df_4b_bottom5)
```

```
        county_name  geotypevalue  total_households  burdened_househ
olds  \
36827      Monterey    6053010306             180.0
0.0
392094       Sutter    6101050402             267.0
0.0
145583  Los Angeles    6037433802             110.0
0.0
389988    Riverside    6065046601              67.0
0.0
388476    Riverside    6065045228             322.0
0.0

        percent
36827       0.0
392094      0.0
145583      0.0
389988      0.0
388476      0.0
```

In [49]:
```
# Calculate mean
df_4b.percent.mean()
```

Out[49]: 25.5340296609756

In [50]:
```
# Calculate median
df_4b.percent.median()
```

Out[50]: 25.41324722

Question 5a

In [51]:
```
# We will create a subset of our dataframe that captures the rows of i
nterest as specifed by 5a.
# 5. In the Bay Area Region (by Census Tract), what percentage of hous
eholds (all race/ethnic groups) are affected by a cost burden >= 50% m
onthly household income?
# What is the mean and median for each?
#       a. Mortgage paying, owner occupied households

df_5a = data_mod4[(data_mod4.region_name == 'Bay Area') & (data_mod4.g
eotype == 'CT') & (data_mod4.burden == '>= 50% of monthly household in
come consumed by monthly, selected housing costs') & (data_mod4.tenure
== 'Mortgage-paying, owner-occupied households') & (data_mod4.income_l
evel == 'All income levels') &  (data_mod4.race_eth_name == 'Total')];
df_5a = df_5a.dropna(subset = ["percent"], inplace=False);
```

In [52]:
```
# Show the 5 Bay Area Census Tracts (minimum of 50 households) with th
e Highest Percentage of Households with the specified burden criteria
df_5a_top5 = df_5a.sort_values(by='percent',ascending = False)
df_5a_top5 = df_5a_top5[df_5a_top5["total_households"]>=50]
df_5a_top5 = df_5a_top5.head(5)
df_5a_top5 = df_5a_top5[['county_name', 'geotypevalue', 'total_househo
lds', 'burdened_households', 'percent']]
print(df_5a_top5)
```

```
          county_name  geotypevalue  total_households  burdened_hous
eholds  \
118736   Contra Costa    6013379000             587.0
321.0
135952        Alameda    6001408800             391.0
213.0
330992  San Francisco    6075012800             603.0
327.0
428643         Solano    6095253300             395.0
212.0
136070        Alameda    6001409000             406.0
217.0

          percent
118736  54.684838
135952  54.475703
330992  54.228856
428643  53.670886
136070  53.448276
```

In [53]:
```
# Show the 5 Bay Area Census Tracts (minimum of 50 households) with th
e Lowest Percentage of Households with the specified burden criteria
df_5a_bottom5 = df_5a.sort_values(by='percent',ascending = True)
df_5a_bottom5 = df_5a_bottom5[df_5a_bottom5["total_households"]>=50]
df_5a_bottom5 = df_5a_bottom5.head(5)
df_5a_bottom5 = df_5a_bottom5[['county_name', 'geotypevalue', 'total_h
ouseholds', 'burdened_households', 'percent']]
print(df_5a_bottom5)
```

```
          county_name  geotypevalue  total_households  burdened_hous
eholds  \
322407  San Francisco    6075011000              82.0
0.0
327374  San Francisco    6075011902              53.0
0.0
112094   Contra Costa    6013328000              50.0
0.0
336780  San Francisco    6075015802             228.0
0.0
129536        Alameda    6001405402             187.0
0.0

        percent
322407      0.0
327374      0.0
112094      0.0
336780      0.0
129536      0.0
```

In [54]:
```
# Calculate mean
df_5a.percent.mean()
```

Out[54]: 21.974280040568836

In [55]:
```
# Calculate median
df_5a.percent.median()
```

Out[55]: 21.175786305000003

Question 5b

In [56]:
```
# We will create a subset of our dataframe that captures the rows of i
nterest as specifed by 5b.
# 5. In the Bay Area Region (by Census Tract), what percentage of hous
eholds (all race/ethnic groups) are affected by a cost burden >= 50% m
onthly household income?
# What is the mean and median for each?
#       b. Rent paying, renter occupied households

df_5b = data_mod4[(data_mod4.region_name == 'Bay Area') & (data_mod4.g
eotype == 'CT') & (data_mod4.burden == '>= 50% of monthly household in
come consumed by monthly, gross rent') & (data_mod4.tenure == 'Rent-pa
ying, renter-occupied households') & (data_mod4.income_level == 'All i
ncome levels') & (data_mod4.race_eth_name == 'Total')];
df_5b = df_5b.dropna(subset = ["percent"], inplace=False);
```

In [57]:
```
# Show the 5 Bay Area Census Tracts (minimum of 50 households) with th
e Highest Percentage of Households with the specified burden criteria
df_5b_top5 = df_5b.sort_values(by='percent',ascending = False)
df_5b_top5 = df_5b_top5[df_5b_top5["total_households"]>=50]
df_5b_top5 = df_5b_top5.head(5)
df_5b_top5 = df_5b_top5[['county_name', 'geotypevalue', 'total_househo
lds', 'burdened_households', 'percent']]
print(df_5b_top5)
```

```
        county_name  geotypevalue  total_households  burdened_househ
olds  \
448686  Santa Clara    6085512510             786.0                4
76.0
132082      Alameda    6001407102             576.0                3
44.0
130787      Alameda    6001406400             317.0                1
84.0
434484  Santa Clara    6085503602             630.0                3
61.0
147149      Alameda    6001436700             458.0                2
60.0

          percent
448686  60.559796
132082  59.722222
130787  58.044164
434484  57.301587
147149  56.768559
```

In [58]:
```
# Show the 5 Bay Area Census Tracts (minimum of 50 households) with th
e Lowest Percentage of Households with the specified burden criteria
df_5b_bottom5 = df_5b.sort_values(by='percent',ascending = True)
df_5b_bottom5 = df_5b_bottom5[df_5b_bottom5["total_households"]>=50]
df_5b_bottom5 = df_5b_bottom5.head(5)
df_5b_bottom5 = df_5b_bottom5[['county_name', 'geotypevalue', 'total_h
ouseholds', 'burdened_households', 'percent']]
print(df_5b_bottom5)
```

```
       county_name  geotypevalue  total_households  burdened_househo
lds  \
453816   San Mateo    6081608023              95.0
0.0
456246   San Mateo    6081611600             139.0
0.0
127331     Alameda    6001404300             171.0
0.0
148823     Alameda    6001440332              73.0
0.0
127708     Alameda    6001404501              94.0
0.0

        percent
453816      0.0
456246      0.0
127331      0.0
148823      0.0
127708      0.0
```

In [59]:
```
# Calculate mean
df_5b.percent.mean()
```

Out[59]: 23.001873349023814

In [60]:
```
# Calculate median
df_5b.percent.median()
```

Out[60]: 22.440654625

Data Visualization

The dataframes above were generated to capture the specific data needed to further investigate and answer the initial questions we had before performing our deep dive into our data. The data collected further guides our exploratory analysis and helps to potentially uncover new relationships within the data or understand the significance of our results. In this section, we will create data visualizations to capture and illustrate the major insights uncovered from our exploratory analysis.

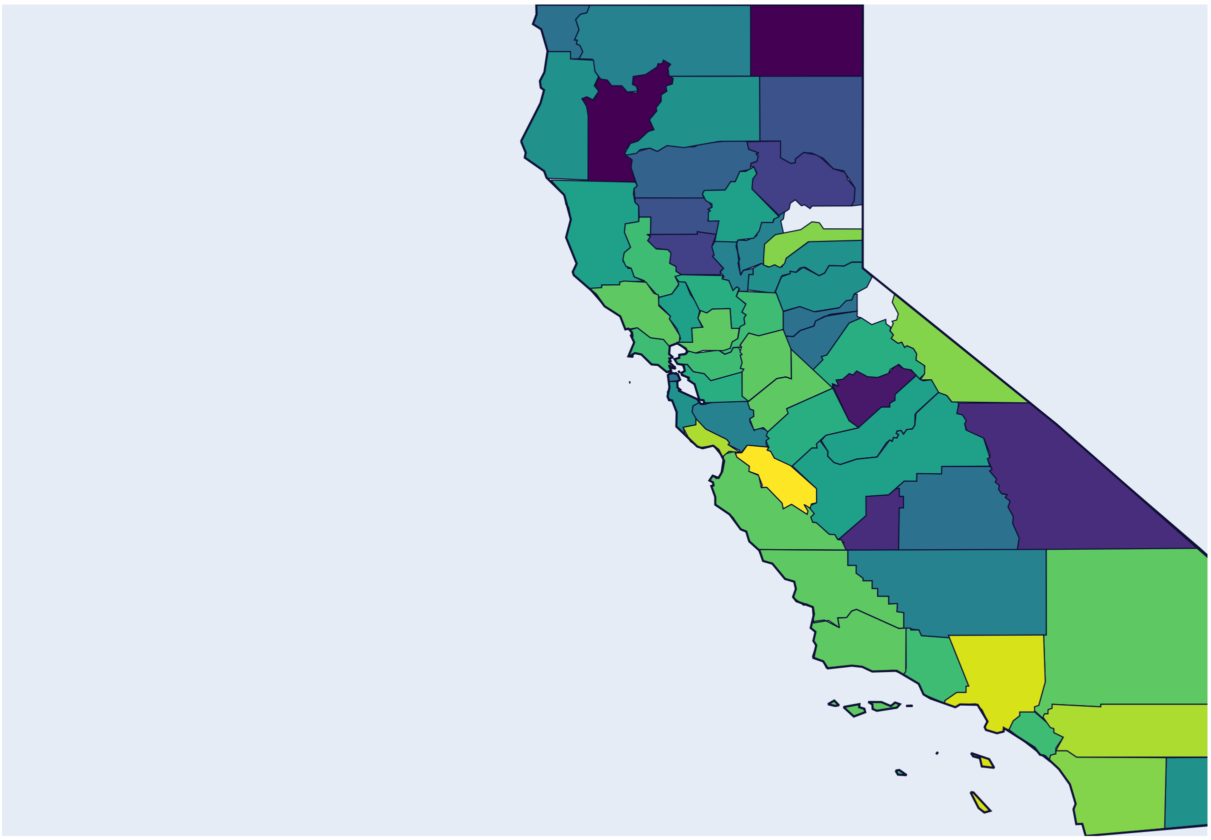
Housing Cost Burdens by Geographic Location (State-wide and Bay Area Region-Wide)

Question 1 aimed to explore the housing cost burden percentages across all California counties. The two severity levels of cost burden that were evaluated were at the " > 30%" cost burden and " > 50%" cost burden, whether that be from rent or selected housing costs/mortgage. Also, for further inclusivity, this analysis considered all race/ethnic groups and accounted for all HUD-adjusted family median income. Looking at our results, the maximum percentage of households experiencing at least a 30% cost burden in each region was approximately 49.8% (San Benito County) while the minimum percentage of households was approximately 28.6% (North Coast). With the mean and median being very close (41.8% and 42.5%, respectively), the distribution of the percentage of households with a 30% cost burden appears to skew closer to the upper end. In part b of this question, we look at another metric at similar conditions except taken at least a 50% cost burden. To no surprise the region that had the maximum percentage of households with that burden (Los Angeles, at 24.5%) and the minimum percentage (Trinity at 10.6%) followed suit. Again, it was observed the mean and median (19.5% and 20.0%) overall skews towards these higher percentages.

For Question 2, we examine the same indicator metrics as Question 1, but instead narrow our scope to just the Bay Area Region and do so with a breakdown at the county-level. Today, the Bay Area is widely known to have some of (if not the) highest housing costs in the entire nation. Currently, residing in the Bay Area myself and having to deal with these high costs, I was curious to see what the housing cost burden looked like and how it varied county by county back between the years of 2006 and 2010. These high housing costs and therefore cost of living, was driven by the influx of well paid tech workers employed to the many large Bay Area tech companies (such Apple, Google, Facebook, etc). The sheer amount of high salary workers looking for homes in the Bay Area overwhelemed the available housing, thus causing a shortage and driving up the prices. Looking at our dataframe, a maximum of 45% of households in Solano County were burdened by a housing costs that were at least 30% of the household's monthly income while the minimum percentage of households burdened by this was in San Francisco with only 39.4%. This distribution and variation doesn't appear to be too drastic across the different counties of the Bay Area region and overall they are relatively close (~3% difference) to the mean and median values calculated for all California counties . This parallels the >50% monthly income burden case but with Marin county at the top with 21.5% and Santa Clara county at the bottom with 18.4%. These max and min values fit the corresponding median and mean values calculated in the previous section (20.2%). Therefore, overall, between the years of 2006-2010, there does not appear to be a large disparity in burdened household percentages between counties in the Bay Area region with state counties as a whole.

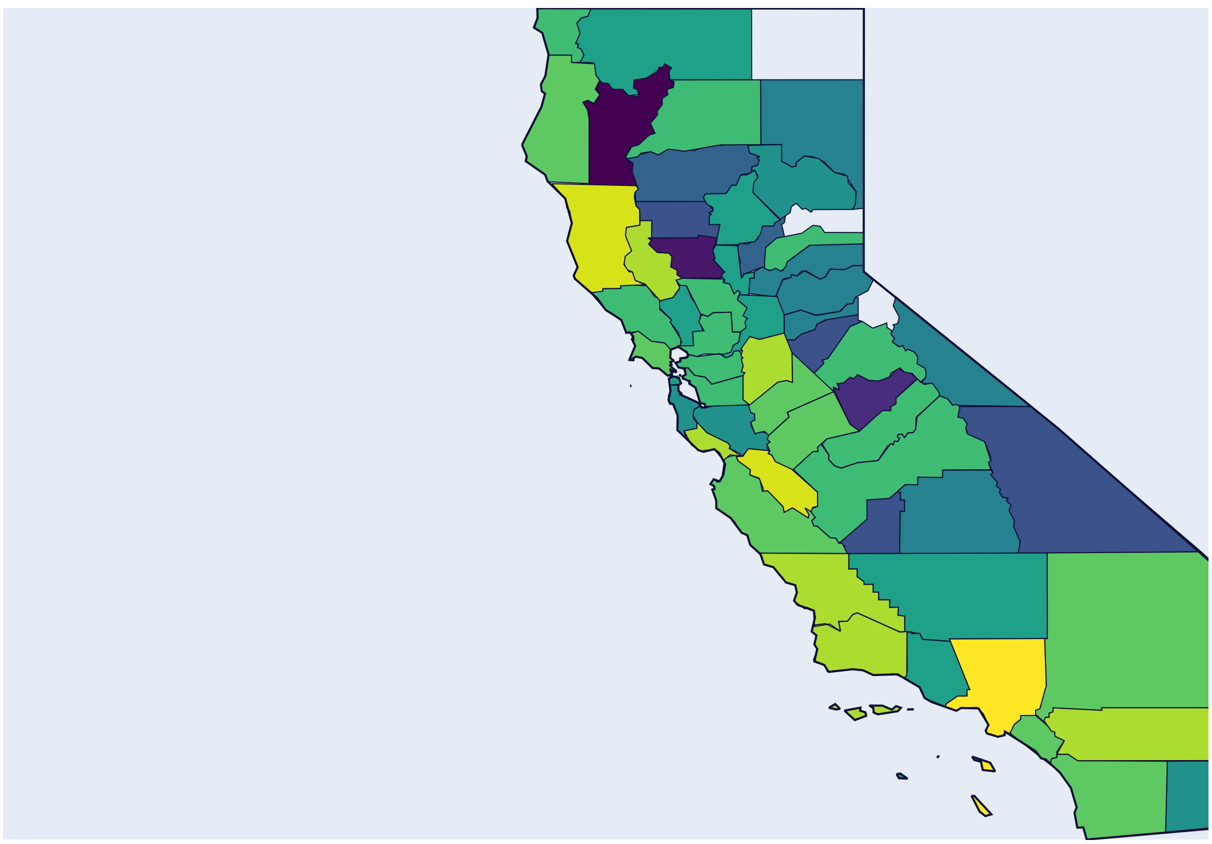
In [61]:
```
# Graph of California counties with cost burdens > 30% monthly income
fig = ff.create_choropleth(
    fips=df_1a['geotypevalue'],
    values=df_1a['percent'].astype(int),
    show_state_data=True,
    scope=['CA'], # Define your scope
    county_outline={'color': 'rgb(15, 15, 55)', 'width': 0.5},
    state_outline={'color': 'rgb(15, 15, 55)', 'width': 1},
    legend_title='Percent', title='California State-wide Housing Cost Burdens > 30% Monthly Housing Income'
)
fig.show()
```

California State-wide Housing Cost Burdens > 30% Monthly Ho



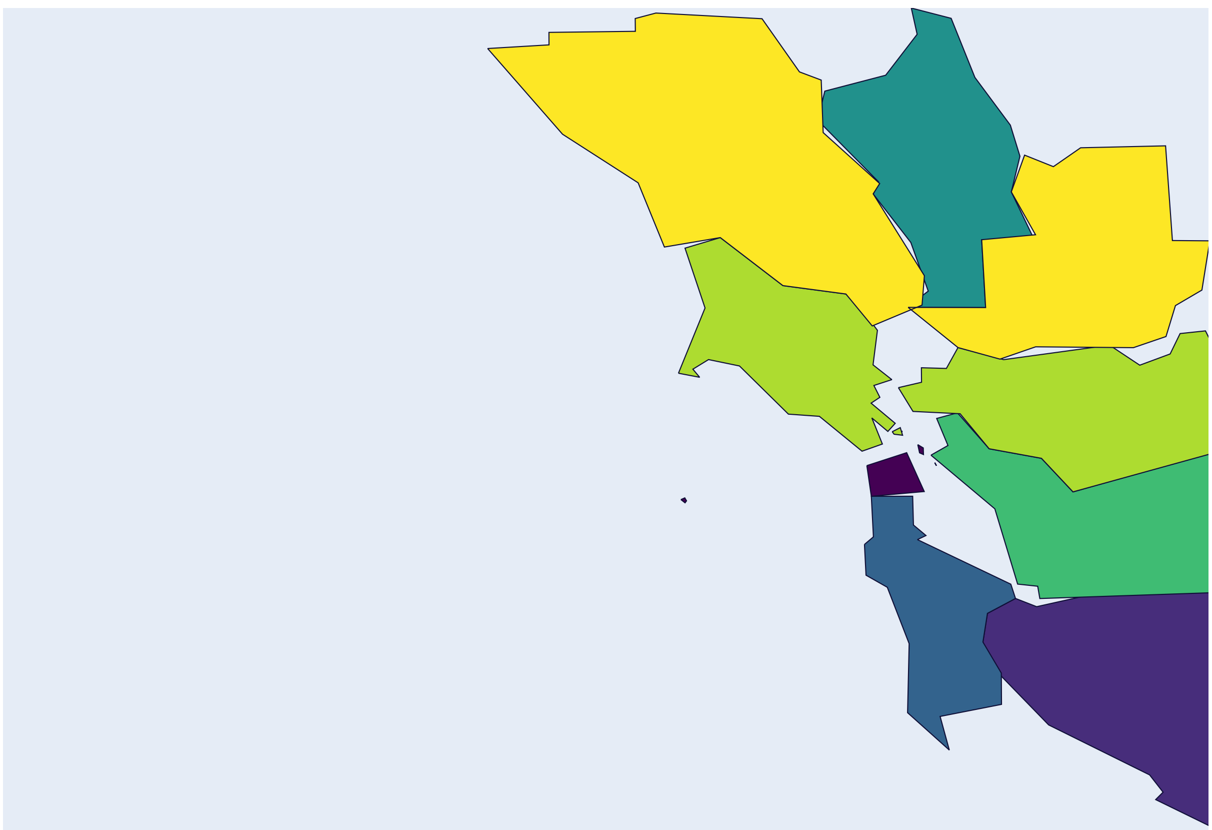
In [62]:
```
# Graph of California counties with cost burdens > 50% monthly income
fig = ff.create_choropleth(
    fips=df_1b['geotypevalue'],
    values=df_1b['percent'].astype(int),
    show_state_data=True,
    scope=['CA'], # Define your scope
    county_outline={'color': 'rgb(15, 15, 55)', 'width': 0.5},
    state_outline={'color': 'rgb(15, 15, 55)', 'width': 1},
    legend_title='Percent', title='California State-wide Housing Cost Burdens > 50% Monthly Housing Income'
)
fig.show()
```

California State-wide Housing Cost Burdens > 50% Monthly Ho



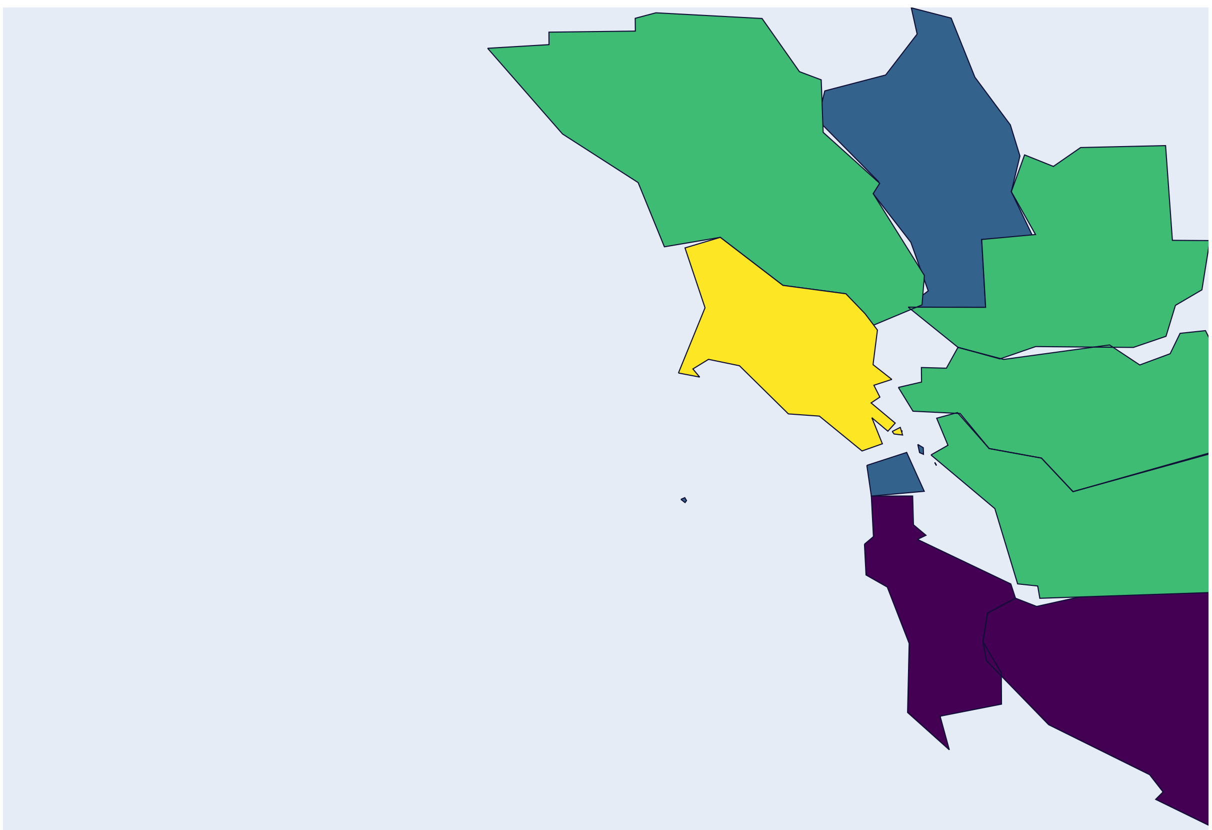
In [63]:
```
# Graph of Bay Area counties with cost burdens > 30% monthly income
fig = ff.create_choropleth(
    fips=df_2a['geotypevalue'],
    values=df_2a['percent'].astype(int),
    show_state_data=True,
    scope=['Bay'],
    county_outline={'color': 'rgb(15, 15, 55)', 'width': 0.5},
    state_outline={'color': 'rgb(15, 15, 55)', 'width': 1},
    legend_title='Percent', title='Bay Area Housing Cost Burdens > 30%
Monthly Household Income'
)
fig.show()
```

Bay Area Housing Cost Burdens > 30% Monthly Household Inc



In [64]:
```
# Graph of Bay Area counties with cost burdens > 50% monthly household
income
fig = ff.create_choropleth(
    fips=df_2b['geotypevalue'],
    values=df_2b['percent'].astype(int),
    show_state_data=True,
    scope=['Bay'], # Define your scope
    county_outline={'color': 'rgb(15, 15, 55)', 'width': 0.5},
    state_outline={'color': 'rgb(15, 15, 55)', 'width': 1},
    legend_title='Percent', title='Bay Area Housing Cost Burdens > 50%
Monthly Household Income'
)
fig.show()
```

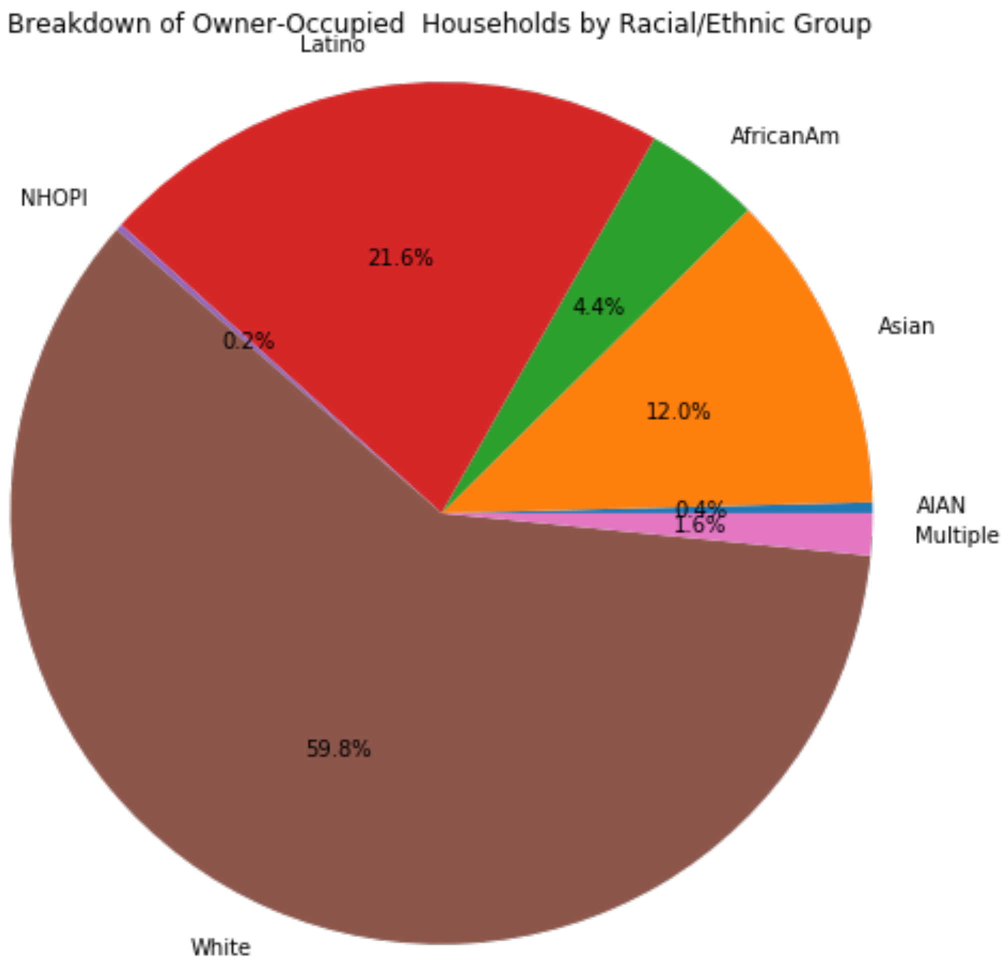
Bay Area Housing Cost Burdens > 50% Monthly Household Inc



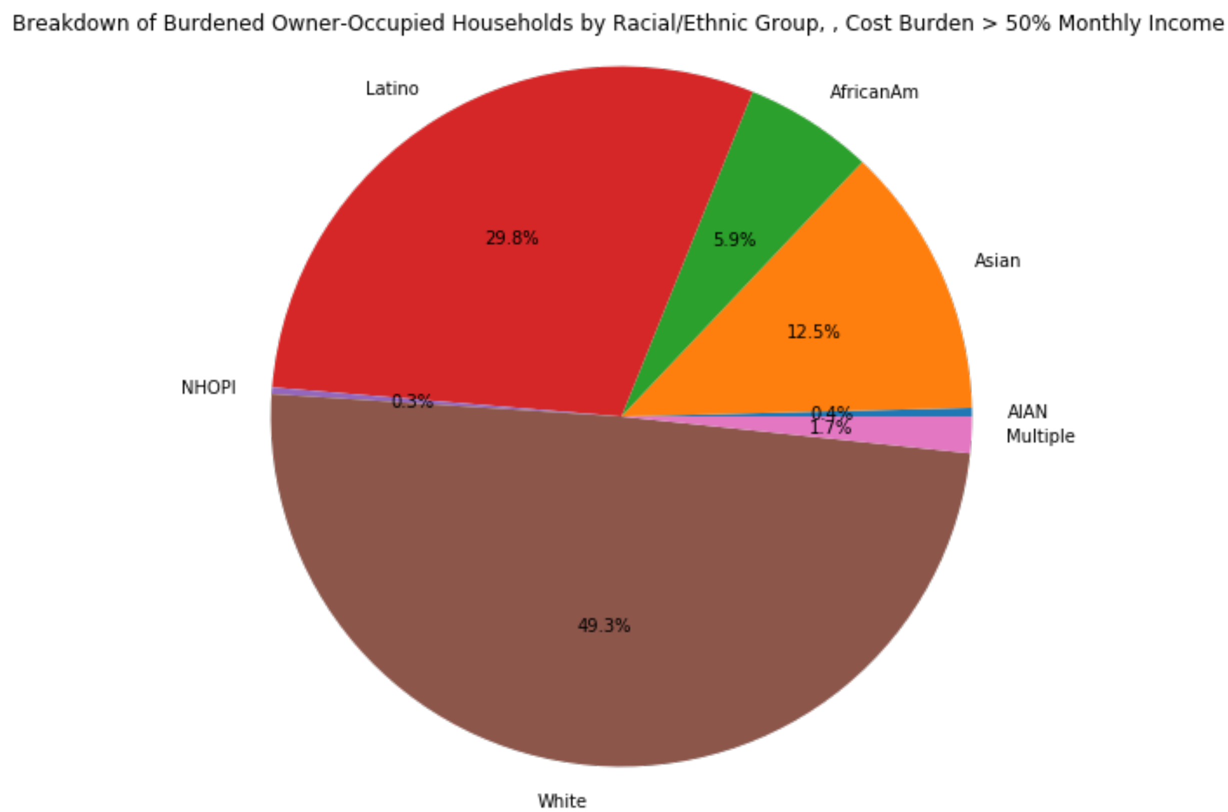
Housing Cost Burdens by Racial/Ethnic Group (State-wide)

The focus of question 3 was to assess how the percentage of burdened households varied by different ethnic/racial groups. In this part, we again take into consideration the entire state of California as a whole regardless of income levels and further breakdown the assessment by splitting it into (oth owner-occupied and renter-occupied households. Our data shows that, with respect to owner-occupied households, Latinos were at the top of both indicators showing household percetanges whose monthly incomes were consumed by at least 30% and 50% by housing costs (51.8% and 21.2%, respectively). On the otherhand, the White demographic comprised of the lowest percentage of households least burdened by at least 30% and 50% of their monthly income going toward housing costs (36% and 15.1%, respectively). With respect to renter-occupied households, a similar disparity exists but for other race/ethnic groups. At 57.9% of African American households spending at least 30% of their monthly income on rent and 32.7% spending at least 50% on rent, the African American demographic is the group that is most burdened by housing costs concerning renter-occupied tenure. Asian households represented the smallest percentage were at least 30% of the monthly income was spent on rent, while only 21.6% of NHOPI households represented the smallest percentage spending at least 50% of income on rent. Compared to the difference between the highest and lowest percentages broken down at a region to region level, the disparity when comparing the highest and lowest percentages of households burdened by the same indicator is significantly larger when we examine the data based on race/ethnic group. This tells us that there is greater housing cost burden inequality at a racial/ethnic demographic level then there is at a geographic level. This is further supported by the large difference when comparing the mean/median values at an overall state-wide level vs. the maximum and minimum values shown at the race/ethnic group level. For instance, where 51.8% of Latino households were spending their 30% of monthly income on housing costs, the corresponding mean and median value (State-wide) were 43.1% and 42.5%, respectively, which is a significant difference.

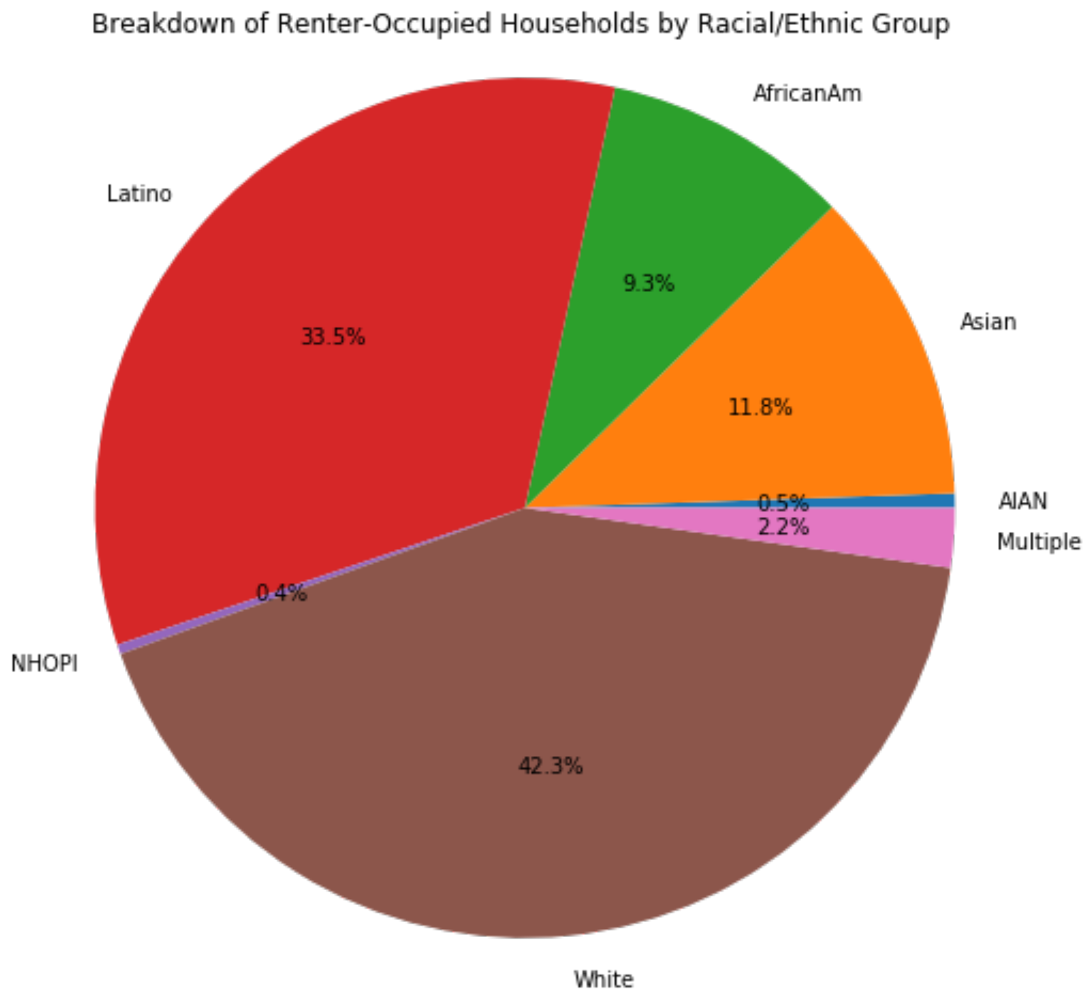
In [65]:
```
labels = df_3b['race_eth_name']
sizes = df_3b['total_households']
plt.rcParams["figure.figsize"] = (8,8)

fig1, ax1 = plt.subplots()
ax1.pie(sizes, labels=labels, autopct='%1.1f%%', shadow=False, startan
gle=0)
plt.title("Breakdown of Owner-Occupied  Households by Racial/Ethnic Gr
oup")
ax1.axis('equal')
plt.show()
```



In [66]:
```
labels = df_3b['race_eth_name']
sizes = df_3b['burdened_households']
plt.rcParams["figure.figsize"] = (8,8)

fig1, ax1 = plt.subplots()
ax1.pie(sizes, labels=labels, autopct='%1.1f%%', shadow=False, startan
gle=0)
plt.title("Breakdown of Burdened Owner-Occupied Households by Racial/E
thnic Group, , Cost Burden > 50% Monthly Income")
ax1.axis('equal')
plt.show()
```

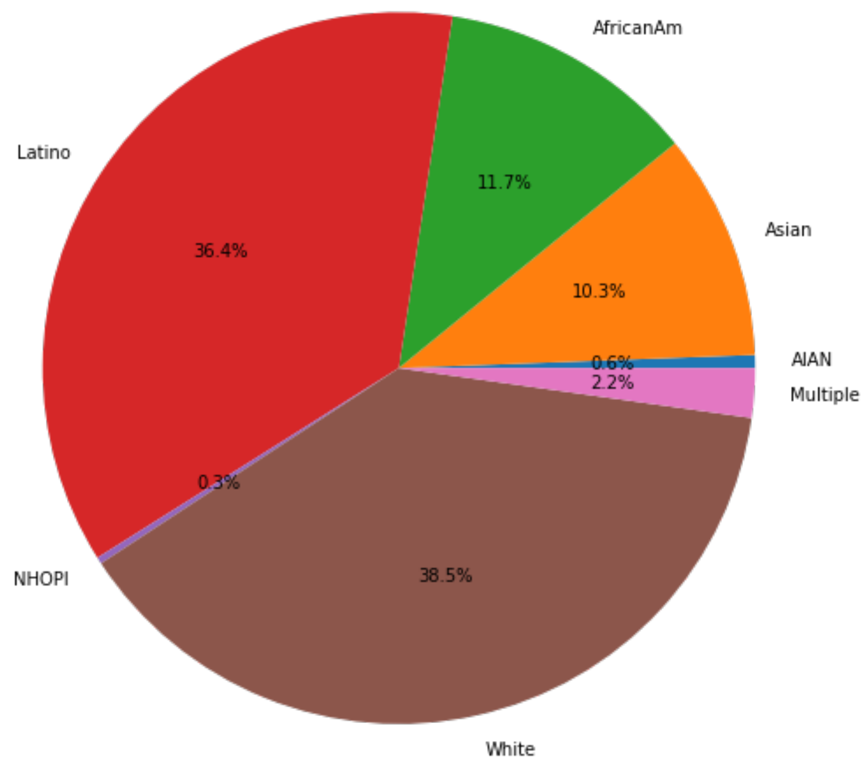


In [67]:
```
labels = df_3d['race_eth_name']
sizes = df_3d['total_households']
plt.rcParams["figure.figsize"] = (8,8)

fig1, ax1 = plt.subplots()
ax1.pie(sizes, labels=labels, autopct='%1.1f%%', shadow=False, startan
gle=0)
plt.title("Breakdown of Renter-Occupied Households by Racial/Ethnic Gr
oup")
ax1.axis('equal')
plt.show()
```



In [68]:
```
labels = df_3d['race_eth_name']
sizes = df_3d['burdened_households']
plt.rcParams["figure.figsize"] = (8,8)


fig1, ax1 = plt.subplots()
ax1.pie(sizes, labels=labels, autopct='%1.1f%%', shadow=False, startan
gle=0)
plt.title("Breakdown of Burdened Renter-Occupied Households by Racial/
Ethnic Group, Cost Burden > 50% Monthly Income")
ax1.axis('equal')
plt.show()
```

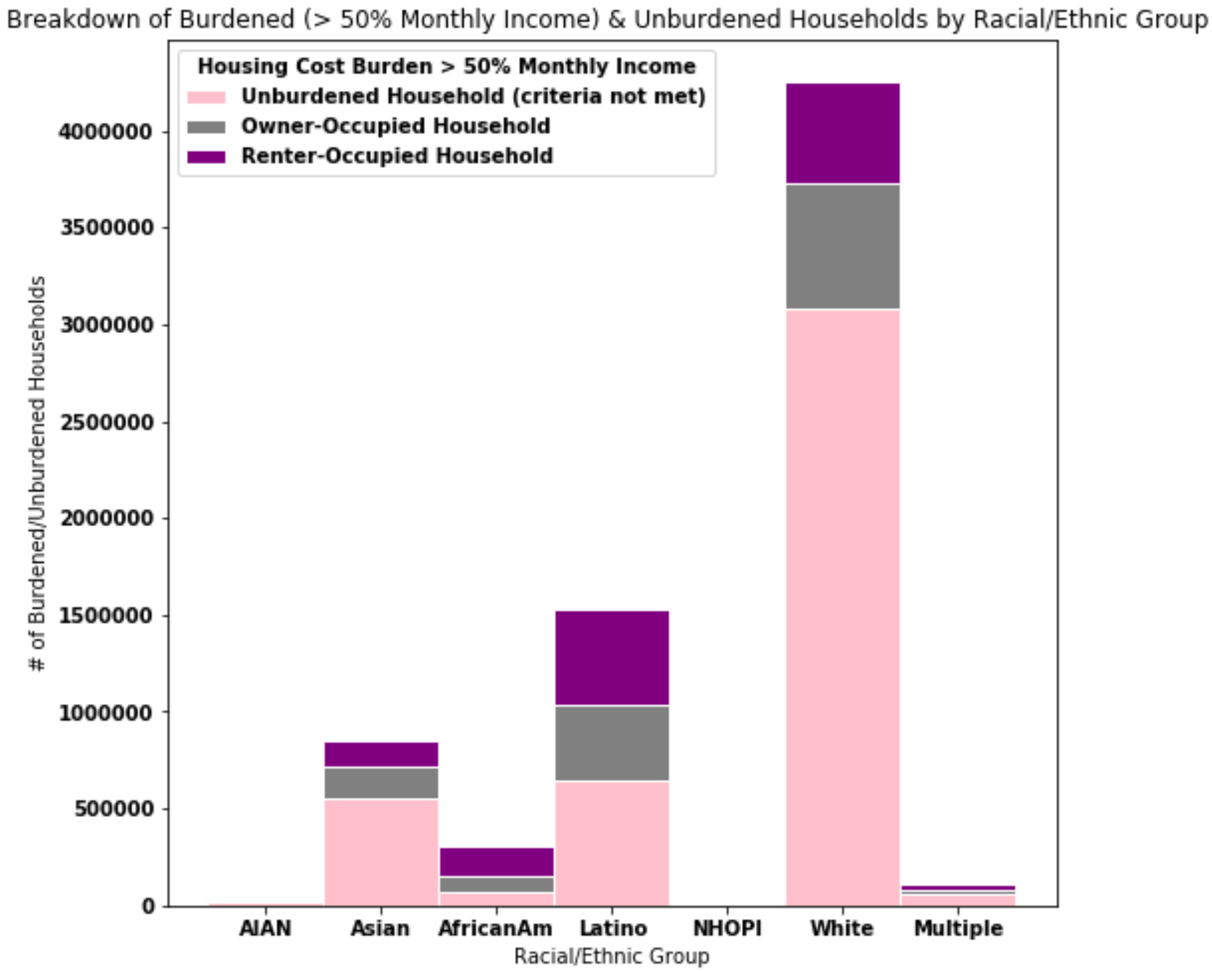
Breakdown of Burdened Renter-Occupied Households by Racial/Ethnic Group, Cost Burden > 50% Monthly Income



In [69]:
```
rc('font', weight='bold')
#Unburdened Households (Subtract owner occupied burdened and renter-oc
cupied burdened from total households)
bars1 = [df_3b.loc[13, 'total_households'] - df_3b.loc[13, 'burdened_h
ouseholds'] - df_3d.loc[15, 'burdened_households'],
         df_3b.loc[19, 'total_households'] - df_3b.loc[19, 'burdened_h
ouseholds'] - df_3d.loc[21, 'burdened_households'],
         df_3b.loc[25, 'total_households'] - df_3b.loc[25, 'burdened_h
ouseholds'] - df_3d.loc[27, 'burdened_households'],
         df_3b.loc[31, 'total_households'] - df_3b.loc[31, 'burdened_h
ouseholds'] - df_3d.loc[33, 'burdened_households'],
         df_3b.loc[37, 'total_households'] - df_3b.loc[37, 'burdened_h
```

```
ouseholds'] - df_3d.loc[39, 'burdened_households'],
         df_3b.loc[43, 'total_households'] - df_3b.loc[43, 'burdened_h
ouseholds'] - df_3d.loc[45, 'burdened_households'],
         df_3b.loc[49, 'total_households'] - df_3b.loc[49, 'burdened_h
ouseholds'] - df_3d.loc[51, 'burdened_households']]
#Burdened Owner-Occupied Households
bars2 = [df_3b.loc[13, 'burdened_households'],
         df_3b.loc[19, 'burdened_households'],
         df_3b.loc[25, 'burdened_households'],
         df_3b.loc[31, 'burdened_households'],
         df_3b.loc[37, 'burdened_households'],
         df_3b.loc[43, 'burdened_households'],
         df_3b.loc[49, 'burdened_households']]
#Burdened Renter-Occupied Households
bars3 = [df_3d.loc[15, 'burdened_households'],
         df_3d.loc[21, 'burdened_households'],
         df_3d.loc[27, 'burdened_households'],
         df_3d.loc[33, 'burdened_households'],
         df_3d.loc[39, 'burdened_households'],
         df_3d.loc[45, 'burdened_households'],
         df_3d.loc[51, 'burdened_households']]

# Heights of bars1 + bars2
bars = np.add(bars1, bars2).tolist()

# The position of the bars on the x-axis
r = [0,1,2,3,4,5,6]

# Names of group and bar width
names = df_3b['race_eth_name']
barWidth = 1

# Create brown bars
plt.bar(r, bars1, color='pink', edgecolor='white', width=barWidth, lab
el = 'Unburdened Household (criteria not met)')
# Create green bars (middle), on top of the first ones
plt.bar(r, bars2, bottom=bars1, color='grey', edgecolor='white', width
=barWidth, label = 'Owner-Occupied Household')
# Create green bars (top)
plt.bar(r, bars3, bottom=bars, color='purple', edgecolor='white', widt
h=barWidth, label = 'Renter-Occupied Household')

plt.title("Breakdown of Burdened (> 50% Monthly Income) & Unburdened H
ouseholds by Racial/Ethnic Group ")
plt.ylabel("# of Burdened/Unburdened Households")
plt.xlabel("Racial/Ethnic Group")
plt.xticks(r, names, fontweight='bold')
plt.legend(loc = "upper left", title = 'Housing Cost Burden > 50% Mont
hly Income')
```

```
# Show graphic
plt.show()
```

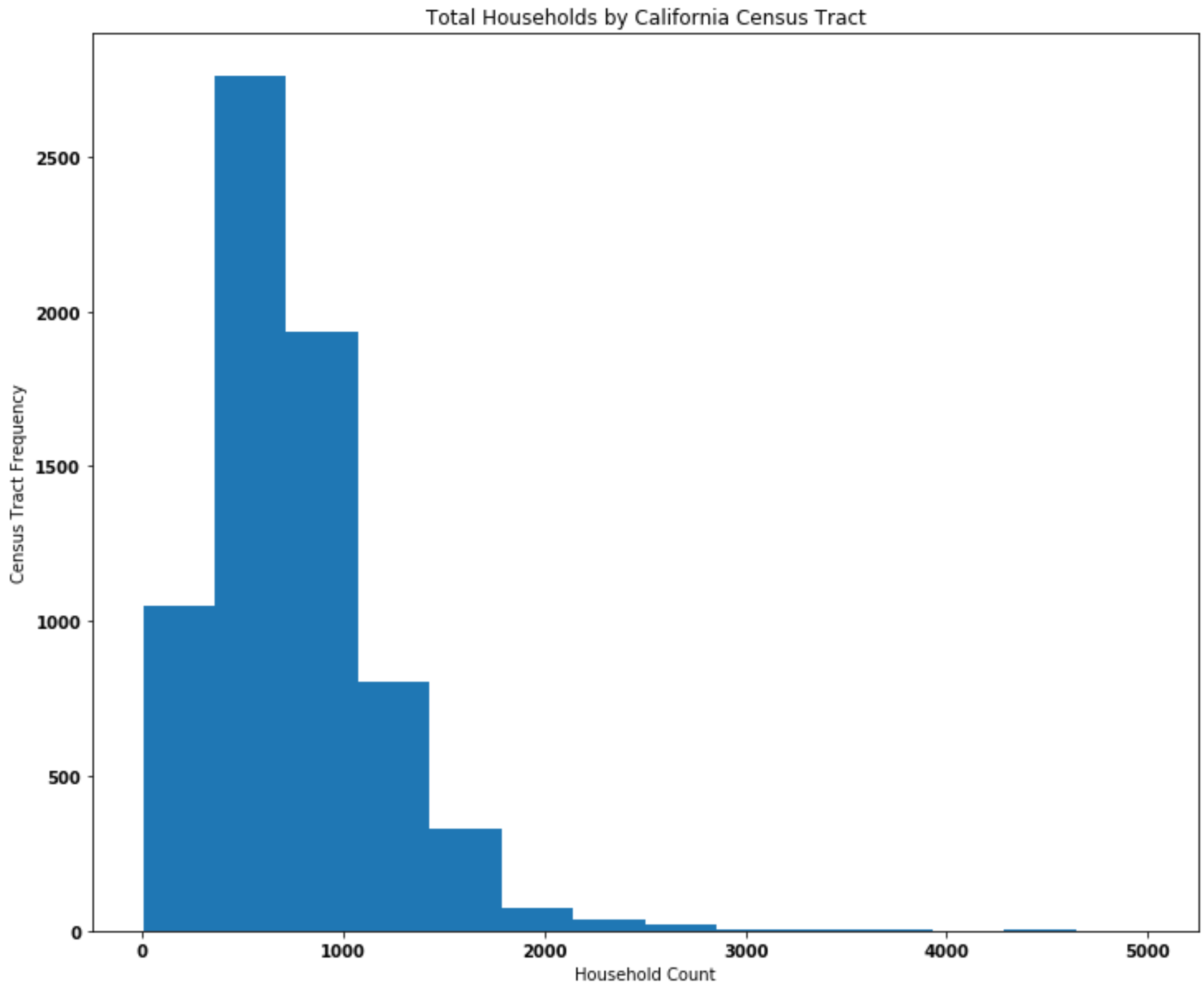


Housing Cost Burdens by Tenure (Mortgage & Rent) - California Census Tracts Statewide

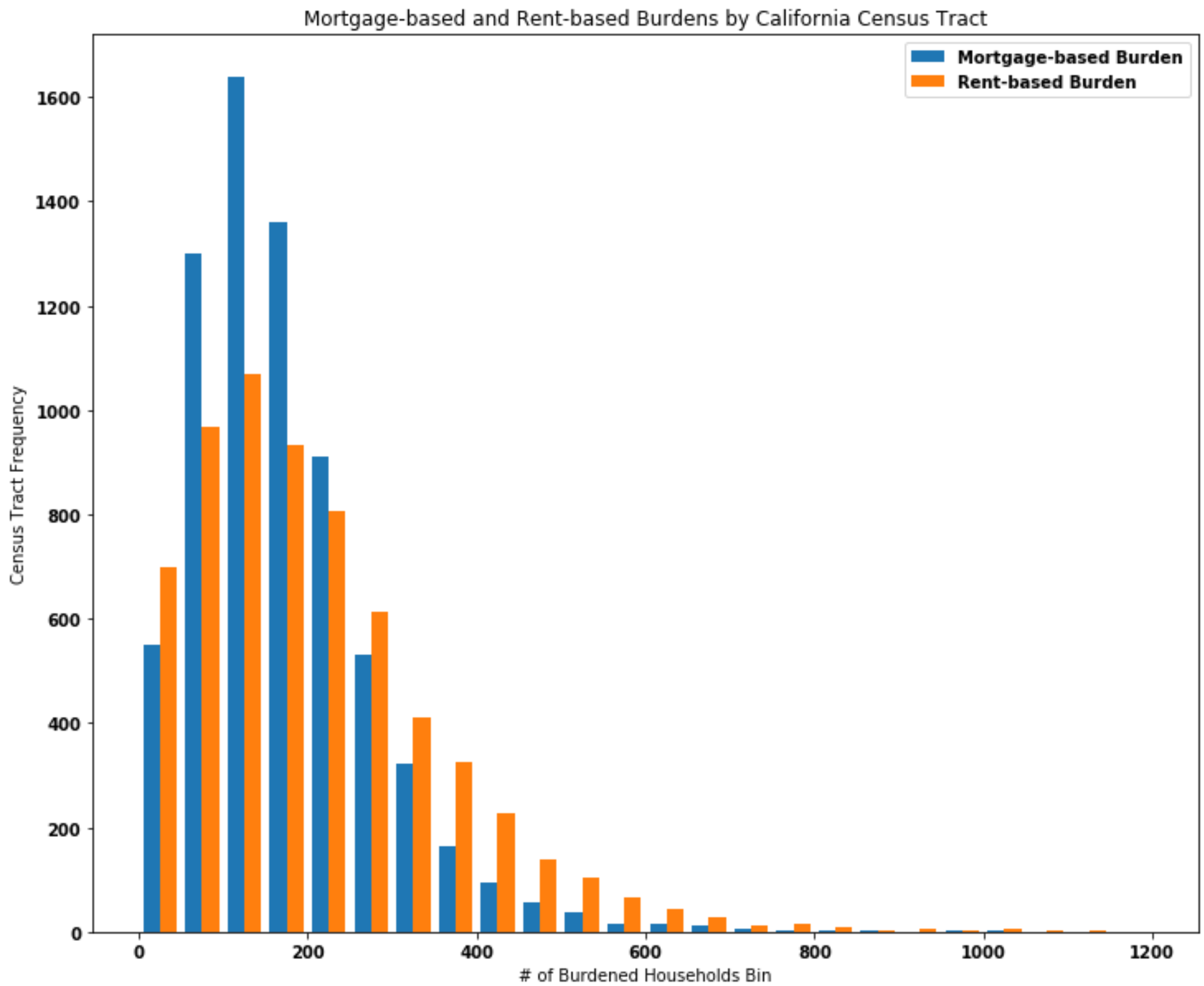
The last two guiding questions in our exploratory analysis address the differences in housing cost burdens Californian households faced solely based on a mortage vs. rent perspective. First, by answering question 4, we delineate difference in tenure at a statewide level by Census Tract (CT). Our results show that of the top 5 burdened (>50% monthly income spent on housing cost) mortgage-paying households, 4 of the census tracts were within Los Angeles county and varied between 56% - 60%. Alarmingly, this is nearly three times the mean/median percentages of households which was calculated to 23.4% and 22.5%. On the otherhand, with respect to this counterpart of rent-paying households, 3 of the top 5 census tracts were in San Luis Obispo County as those results varied between 62% and 75%. Even for renter tenures only, the gap between the top results and the mean/median calculated value (approximately 25.5% for both) is considerably large. This characterizes how skewed the distribution of cost burdened households in California is when breaking our data down by CT. In other words, there are a much larger amount of CTs that show less households being burdened by high cost housing than there are of CTs where high burden costs are more of a problem. This however, can be a misleading statistic as it does not weigh into account the sheer amount of households in each CT and population density (so long as the CT had greater than 50 households in our analysis). However, this indicator and metric is useful in telling us about the concentration of highly burdened households by geographic location.

Lastly, for my own personal interests, we narrow the boundaries and scope of our burden analysis by tenure type to just the Census Tracts within the Bay Area. Our results show how 4 different counties are represented in the top 5 Bay Area Census Tracts with the highest percentages of households with greater than 50% of their monthly income going to mortgage housing costs. In descending order and ranging from 57.2% to 54.5%, these counties are Santa Clara, Contra Costa, San Mateo and Alameda. The calculated mean and median burden percentages for the Bay Area Census tracts are 20.2% and 19.5%, which closely mirrors the disparity we observed with our results at the state-wide level. The higher quantity of census tracts where there is a small percentage or even 0% of burdened households compared to the smaller number census tracts with high percentage (but also significantly more households) skew our results. Compared to mortgage cost burdens, the rental cost burdens appears to be more common among Bay Area census tracts, with the top 5 results ranging from 60.5% down to 57.2%. Santa Clara, Alameda and Sonoma county represent these top 5 census tracts. For comparison, the mean and median values are 20.5% and 20% respectively. Mortgage and rent cost burdens are an important distinction to analyze seperately since there can be differences in the housing needs (long vs. short term, # of individuals within your household) as well as time-based living situation (whether one has lived in the area for a long time as a homeowner or recently moved there for a new job) that can affect the likelihood and the degree/severity of the cost burden.

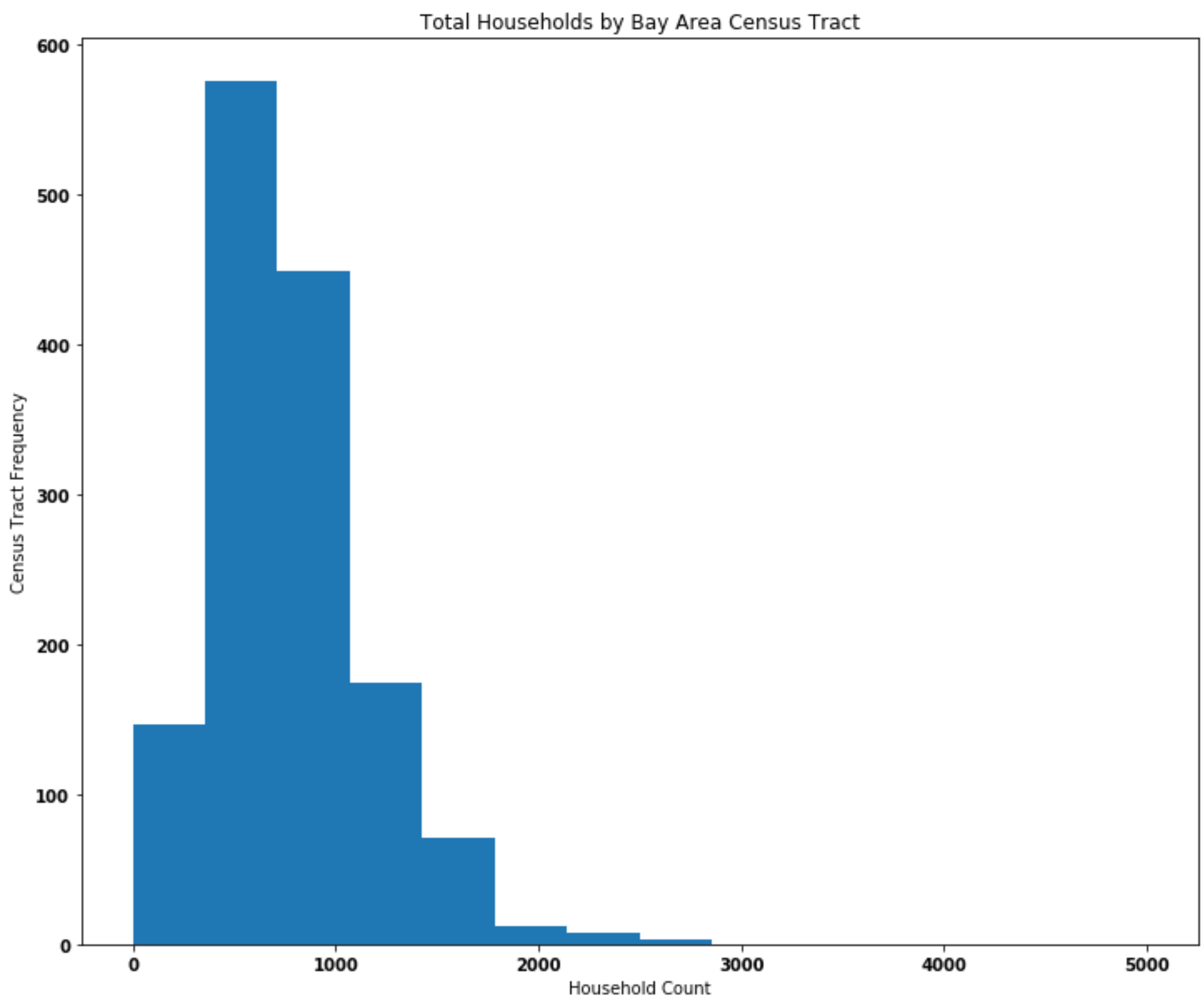
In [70]:
```
# Plot histogram that shows the count of total households across Calif
ornia census tracts
x = df_4a['total_households']
bins = np.linspace(1, 5000, 15)
plt.rcParams["figure.figsize"] = (12,10)

plt.hist([x], bins)
plt.title("Total Households by California Census Tract")
plt.xlabel("Household Count")
plt.ylabel("Census Tract Frequency")
plt.show()
```



In [71]:
```
# Plot histogram that shows the count of mortgage-based and rent-based burdens across California census tracts
x = df_4a['burdened_households']
y = df_4b['burdened_households']
bins = np.linspace(0, 1200, 25)
plt.rcParams["figure.figsize"] = (12,10)

plt.hist([x, y], bins, label=['Mortgage-based Burden', 'Rent-based Burden'])
plt.title("Mortgage-based and Rent-based Burdens by California Census Tract")
plt.legend(loc='upper right')
plt.xlabel("# of Burdened Households Bin")
plt.ylabel("Census Tract Frequency")
plt.show()
```

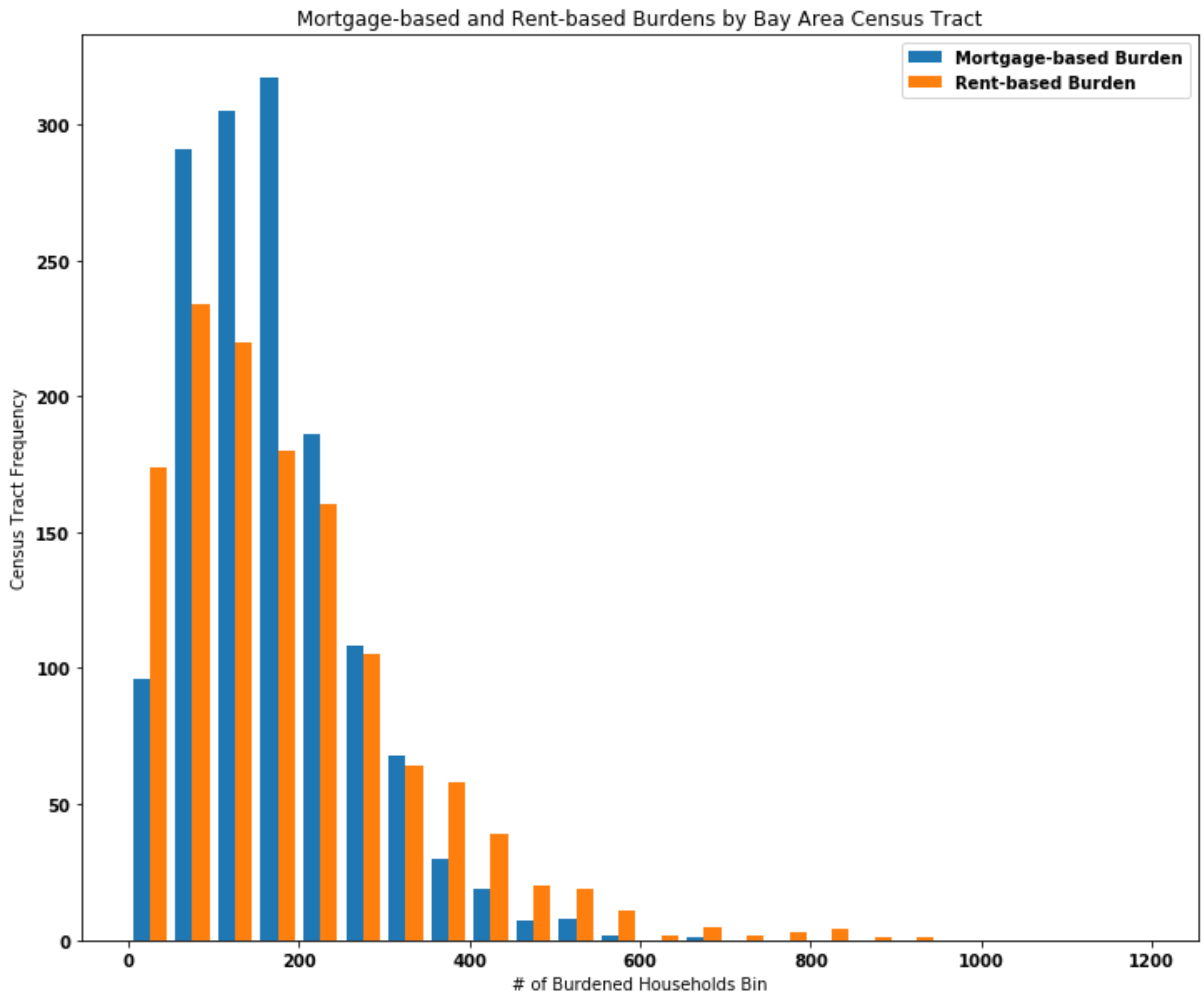


In [72]:
```
# Plot histogram that shows the count of total households across Bay A
rea census tracts
x = df_5a['total_households']
bins = np.linspace(0, 5000, 15)
plt.rcParams["figure.figsize"] = (12,10)

plt.hist([x], bins)
plt.title("Total Households by Bay Area Census Tract")
plt.xlabel("Household Count")
plt.ylabel("Census Tract Frequency")
plt.show()
```



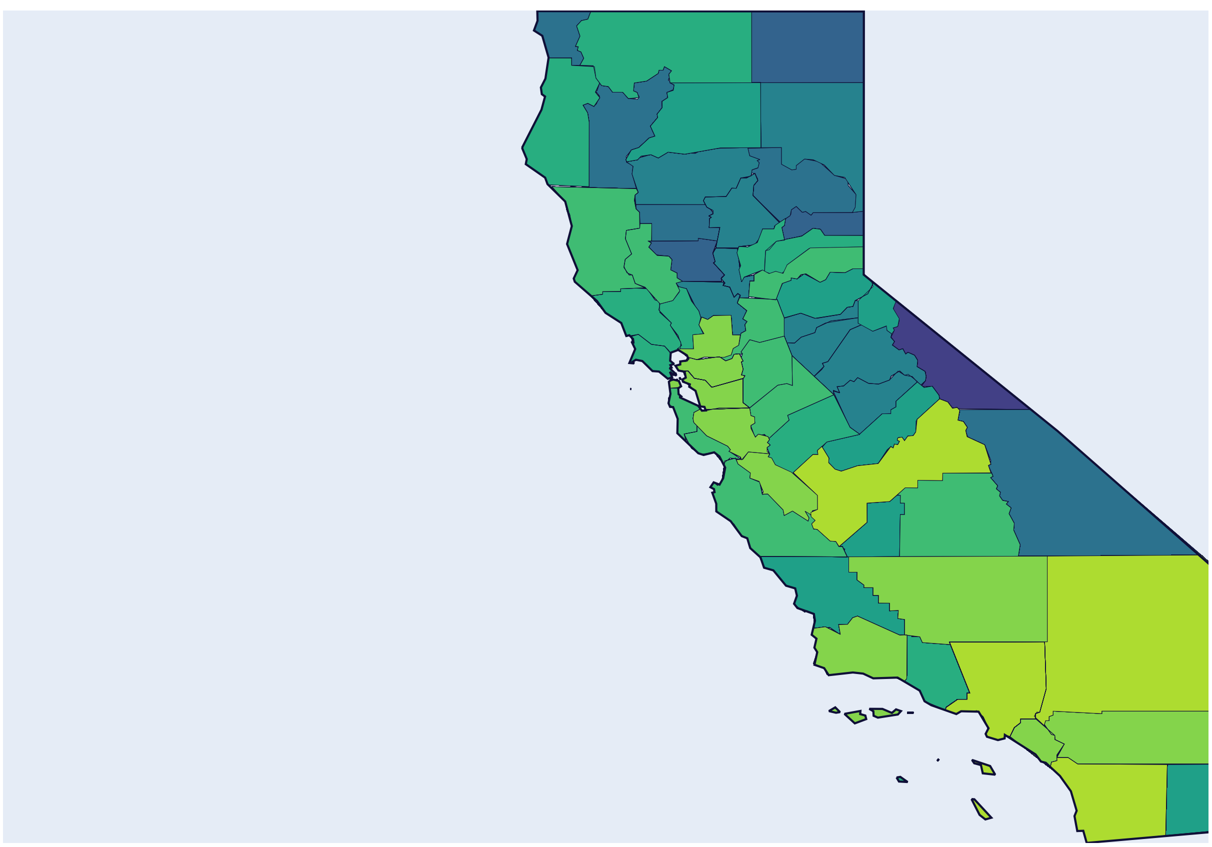
In [73]:
```
#### Housing Cost Burdens by Tenure (Mortgage & Rent) - Bay Area Censu
s Tracts (Region specific)
```

In [74]:
```
# Plot histogram that shows the count of mortgage-based and rent-based
burdens across Bay Area census tracts
x = df_5a['burdened_households']
y = df_5b['burdened_households']
bins = np.linspace(0, 1200, 25)
plt.rcParams["figure.figsize"] = (12,10)

plt.hist([x, y], bins, label=['Mortgage-based Burden', 'Rent-based Bur
den'])
plt.title("Mortgage-based and Rent-based Burdens by Bay Area Census Tr
act")
plt.legend(loc='upper right')
plt.xlabel("# of Burdened Households Bin")
plt.ylabel("Census Tract Frequency")
plt.show()
```



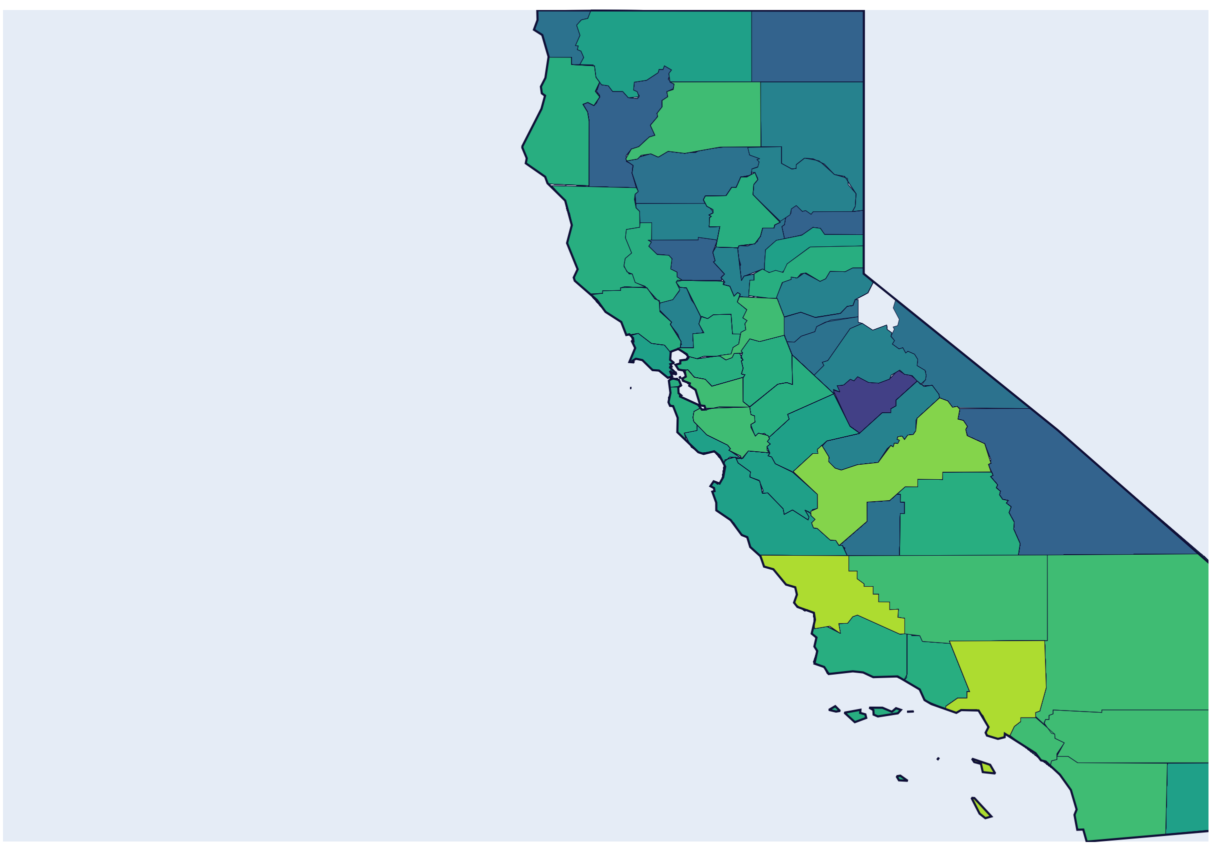
In [78]:
```
# Remove counties with less thatn 50 households as they significantly skew map
df_4a_mod1 = df_4a.drop(df_4a[df_4a.total_households< 50].index)

# Plot Statewide Mortgage-based cost burdens > 50% monthly household income, by county FIPS

endpts = list(np.linspace(1, 65, len(colorscale) - 1))
fig = ff.create_choropleth(
    fips=df_4a_mod1['county_fips'],
    values=df_4a_mod1['percent'].astype(int),
    scope= ['CA'],
    binning_endpoints=endpts,
    county_outline={'color': 'rgb(15, 15, 55)', 'width': 0.001},
    state_outline={'color': 'rgb(15, 15, 55)', 'width': 1},
    legend_title='Percent', title='Statewide Mortgage-based Cost Burdens > 50% Monthly Income (by County FIPS)'
)
fig.show()
```

Statewide Mortgage-based Cost Burdens > 50% Monthly Incon



In [79]:
```
# Remove counties with less thatn 50 households as they significantly
skew map
df_4b_mod1 = df_4b.drop(df_4b[df_4b.total_households< 50].index)

# Plot Statewide Rent-based cost burdens > 50% monthly household incom
e, by county FIPS
endpts = list(np.linspace(1, 81, len(colorscale) -1 ))
fig = ff.create_choropleth(
    fips=df_4b_mod1['county_fips'],
    values=df_4b_mod1['percent'].astype(int),
    scope= ['CA'],
    binning_endpoints=endpts,
    county_outline={'color': 'rgb(15, 15, 55)', 'width': 0.001},
    state_outline={'color': 'rgb(15, 15, 55)', 'width': 1},
    legend_title='Percent', title='Statewide Rent-based Cost Burdens >
50% Monthly Income (by County FIPS)'
)
fig.show()
```

Statewide Rent-based Cost Burdens > 50% Monthly Income (b



Conclusion

The observed insights from the results of our analysis and as illustrated by our data visualizations of the various indicators for Californian Household cost burdents between the years of 2006 and 2010 are summarized as follows:

- Between 2006 and 2010, there is not a drastic difference between the maximum and minimum percentage of burdened households for Bay Area counties compared to the rest of the counties of the state as a whole. However, the overall mean/median burden percentages were skewed toward the higher end as opposed to being close to the midpoint between min/max. This indicates that most counties (especially those in the Bay Area) are "clustered" with having higher burdened household percentages.
- There is an indisputable disparity in the proportionality between unburdened vs. burdened households by racial/ethnic group. The proportionality between Latino and African American households experiencing housing cost burdens compared to their corresponding unburdened households of the same demographic are far higher than White households. 1 in 4 White households are burdened by housing costs while approximately 1 in 2 Latino Households and 2 in 3 African American are burdened.
- From a statewide perspective, there is a correlation between having more mortage-based housing cost burdens than rent-based housing cost burdens with a higher total number of households in census tracts. A lower total number of households in a census tract correlates to a lower frequency of mortage-based cost burdens relative to rent-based ones. The trend holds true when limiting the scope to just households in census tracts of a specific region (e.g. Bay Area). This suggests where housing is more abundant, mortgage owners are more burdened than renters.

Future Work

- Obtain more recent or current data and evaluate trends since the analyzed 2006-2010 time frame.
- Correlate data with mean/median income to see if salaries are being properly adjusted to fit cost of living, high housing costs in specific areas known to have highly burdened households percentages

```
In [ ]:
```